Journal of the
OPTICAL SOCIETY
of AMERICA

VOLUME 35, NUMBER 1 JANUARY, 1945

Colorimeters and Color Standards*

COMMITTEE ON COLORIMETRY

INSTRUMENTS designed for the direct measurement of color in psychophysical terms are called colorimeters. The use of colorimeters for routine measurements has been largely superseded in this country by indirect colorimetry, i.e., the computation of colorimetric specifications from spectrophotometric and certain psychophysical data. The psychophysical data most commonly employed, which were standardized by the International Commission on Illumination, were necessarily acquired through the use of colorimeters. Colorimeters are fundamental for psychophysical investigations of color, and some forms of colorimeters are convenient for the routine measurement and control of restricted classes of colors.[801]

The complications of color measurement are avoided by the use of color standards. These are usually members of a set of numerous and systematically arranged samples which are convenient for making visual comparisons with miscellaneous materials, and for which color specifications are known or obtainable. Color standards are useful for many purposes for which the availability of samples for immediate visual inspection, trial, and comparison is more important than accuracy of specification. The calibrations of some colorimeters are based on color standards, in the form of filters or opaque reflecting surfaces. If the standards have been calibrated in terms of the I.C.I. system, specifications obtained by the use of colorimeters and color standards can be converted to the I.C.I. system. Errors may arise at each intermediate stage of such indirect determinations of I.C.I. specifications, yet convenience and relatively low cost make such methods satisfactory for many problems that do not require a high degree of accuracy.

This chapter is concerned with the description of colorimeters, systems of color standards, and methods of calibration suitable for the expression of their results in terms of the standard I.C.I. system. No specific calibration data are given but reference is made to sources of available data.

1. COLOR COMPARATORS FOR CHEMICAL ANALYSIS

Two classes of instruments, called colorimeters, should be distinguished. One class is employed for the measurement of color with emphasis on the visual equivalence of the sample to a synthesized stimulus. The physical specifications of the components of the synthesized stimulus provide scales for the psychophysical specification of the visually equivalent color. This chapter is concerned primarily with colorimeters designed and used for the specification of color in such terms.

* Chapter VIII of the report of the Committee on Colorimetry of the Optical Society of America. Previously published chapters have been Chapter II (including references 201–208), J. Opt. Soc. Am. **33**, 544–554 (1943); Chapter V (including references 501–573), J. Opt. Soc. Am. **34**, 183–218 (1944); Chapter VI (including references 601–632), J. Opt. Soc. Am. **34**, 245–266 (1944); and Chapter VII (including references 701–760), J. Opt. Soc. Am. **34**, 633–688 (1944).

[801] J. Guild, Proceedings of the Optical Convention (London, 1926), Part 1.

BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

The other class of instruments called colorimeters† is employed in chemical analysis for the determination of concentrations of solutions or suspensions,[802, 803] or for the empirical grading of products according to color. In these instruments, color is employed as an indication of chemical constitution, and the sample is specified in terms of concentration of standard solutions, or in terms of composition of standard materials.[804] Although instruments of this type are commonly called colorimeters, a preferable name is color comparators. They are not designed for the measurement of color but for use in the determination of chemical composition or concentration by color comparisons. Detailed discussions of the construction and use of instruments of this sort will be found elsewhere.[802, 803] Figure 1 is a schematic representation of a typical color comparator employed in chemical analysis. Light from an extended light source *A* (for example, an electric lamp combined with several glass diffusing plates) passes through a cell containing the solution to be tested *B*. Light from the same source also passes through a standard material, usually a solution having a standard composition and concentration and contained in a cell *C*, which maintains a standard depth of the solution in the light path. (Standard light filters of glass or gelatin are sometimes used in place of the standard solution. Such filters are

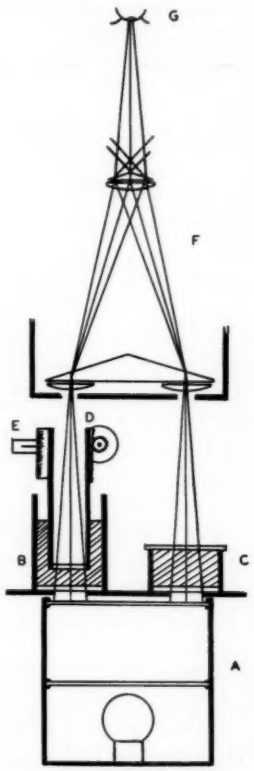


FIG. 1. Schematic diagram of a typical color comparator employed in chemical analysis.

† The names "colorimeter" and "colorimetry" were invented by chemists and have been used for many years with reference to chemical analysis by the use of color comparators. This usage originated before the measurement of color became practical. The first practical instrument for the measurement of color in psychophysical terms was called a colorimeter by its inventor, F. E. Ives. The Colorimetry Section of the United States National Bureau of Standards was created, before 1920, to deal with the measurement of color. These later uses of the names colorimeter and colorimetry are more comparable with the structure of the words than the older usage. Indiscriminate use of these names for both the measurement of color and chemical analysis by color comparisons is responsible for extensive and serious confusion in all libraries and bibliographies of both of these distinctly different sciences. The exclusive use of "colorimeter" and "colorimetry" to refer to instruments and methods for the measurement and specification of color is strongly recommended. The abandonment of the older usages with reference to chemical analysis and grading of materials by color comparisons is urged, with due acknowledgment of the historical priority and great extent of this usage. "Color comparators" is suggested as an appropriate name for the instruments employed for chemical analysis and the grading of the quality of materials by comparisons with color standards. A collective name for the methods employed is not so readily available, but the importance of maintaining a clear distinction between two entirely different sciences warrants careful consideration, and even the coining of a new and more appropriate name. If a clear distinction can be made, in such a manner, by contemporary workers in these two sciences, librarians and bibliographers will undoubtedly adopt the new usage and gladly disentangle the literatures which are now grouped almost indiscriminately under the single subject heading, "Colorimetry." The literatures of both sciences are currently expanding at such a rate that separation of the material will soon become impractical if the establishment of distinctions of the kind suggested is neglected much longer. "Colorimeter" and "colorimetry" will be used throughout this report with exclusive reference to instruments and methods for the measurement of color. This opportunity is taken, at the beginning of the discussion of colorimeters, to mention color comparators and to disclaim any intention of discussing the methods of chemical analysis employing such instruments.

[802] John H. Yoe, *Photometric Chemical Analysis (Colorimetry and Nephelometry)*, (John Wiley & Sons, Inc., New York, 1928–29), two volumes.

[803] Foster D. Snell and Cornelia T. Snell, *Colorimetric Methods of Analysis* (D. Van Nostrand Company, Inc., New York, 1936), two volumes.

[804] H. V. Arny, Proc. 8th Int. Cong. App. Chem. **26**, 319 (1912); H. V. Arny and C. H. Ring, J. Frank. Inst. **180**, 199 (1915); H. V. Arny and C. H. Ring, J. Ind. Eng. Chem. **8**, 309 (1916).

more convenient but they should be carefully designed, constructed, and calibrated for each analytic problem.) The depth of the sample solution B in the light path can be varied by some suitable mechanical-optical device D, such as a hollow cylinder with a transparent base, which can be immersed to variable depths in the solution. The depth of the sample solution in the light path is indicated on a suitable scale E. After passing through the sample B and the standard C the light beams are passed through a suitable optical system F, such as the combination of two biprisms and objective and field lenses illustrated. The observer can then compare in two adjacent areas the light transmitted by the sample and the standard. The depth is adjusted by manipulation of the depth gauge D until the colors in the two adjacent fields match. The concentration of the sample is inversely proportional to the depth (indicated by the scale E) required to produce a color match. Auxiliary light filters are sometimes used in front of the eye of the observer in order to increase the noticeability of the color differences caused by small differences of concentration. The use of such filters in color comparators for chemical analysis is advantageous because interest is confined to the composition of the sample, and the relation of the color of the sample to the color observed in the instrument is unimportant.

2. THREE-COLOR COLORIMETERS USING SPECTRUM COMPONENTS

Many colorimeters have been constructed in which colors are matched by mixtures of three colors of invariable chromaticity but variable luminance. The "colour patch" apparatus used by Abney[805] appears to have been the earliest of this type. In it three spectral colors were isolated by three slits placed in the plane of the spectrum. The luminances of the spectrum components were varied by altering the widths of the slits. Accuracy of measurement was seriously limited by the difficulties of accurate control and calibration of the slit widths. Variation of slit widths also results in changes of chromaticity of the components of the trichromatic mixtures. Although such variations of chromaticity were probably not serious in comparison with the other difficulties and uncertainties encountered with Abney's instrument, such variations would be important sources of error in a colorimeter intended for work comparable in accuracy with modern methods. Compensations for variations of this type would be difficult and of uncertain efficacy.

Allen[806] described a trichromatic colorimeter in which spectra from three different collimators were superimposed in the focal plane of an eyepiece, in juxtaposition with a test color. The field of view was limited to a narrow spectral region by diaphragms in the eyepiece. The luminances were varied by adjusting the widths of the entrance slits of the three monochromators, or by means of calibrated wedges placed in front of the slits. The variation of wave-length in each of the superimposed spectra is a serious limitation on the accuracy of such an instrument.

Wright used a trichromatic colorimeter, Fig. 2,* employing spectrum components, in his redetermination of the color mixture functions (an important part of the experimental work upon which the adoption of the standard I.C.I. color mixture data was based).[614] Wright placed two small totally reflecting prisms, $T.C.$ and $D.P.$, in the plane of the spectrum W_1 formed by the upper monochromator prisms A–A. The prisms $T.C.$ and $D.P.$ served as two slits for the lower monochromator which further purified the reflected spectrum components. After reflection in the prism D these components were brought to a common focus on the pupil of the eye of the observer E. The prism R_1 split the beam from the upper monochromator so that a second spectrum was formed at W_2. In this spectrum were located three small totally reflecting prisms R, G, B. The spectrum components reflected by these prisms were redirected by the prism R_2 and further purified by refraction in the lower monochromator. After reflection in the prisms D and P, these components were brought to a common focus on the pupil of the eye of the observer E. The mixtures of the two sets of components reflected from the spectra W_1 and W_2 were seen as filling adjacent areas of the field

[805] Sir W. deW. Abney, *Colour Measurement and Mixture* (London, 1891).

[806] F. Allen, J. Opt. Soc. Am. **8**, 339 (1924).

* Reprinted by permission of W. D. Wright.

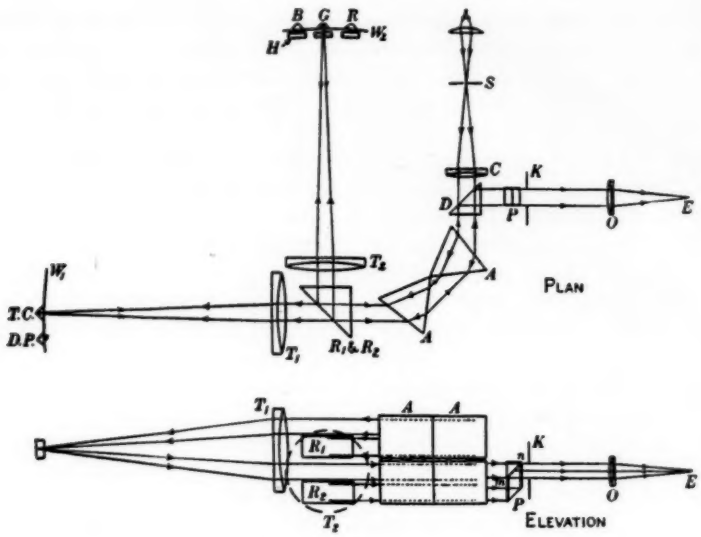


FIG. 2. Wright's trichromatic colorimeter.

prism *P*, forming an excellent photometric field. The prisms *R*, *G*, *B* reflected the three fixed wave-lengths, 650 mμ, 530 mμ, and 460 mμ, which Wright used as primaries for his color-mixture experiments. Calibrated wedges *H* were placed in front of each of these prisms for the control of the intensities of the primaries. The prism *T.C.* reflected the portion of the spectrum for which the color-mixture data were to be determined in any instance. The prism *D.P.* reflected whichever primary was necessary to desaturate the test color so that it could be matched by a mixture of the other two primaries. Wright's trichromatic colorimeter proved highly satisfactory for the determination of color-mixture data, and is unusually well designed for research on the visual functions. It has not been used for ordinary measurements of color, and it is doubtful whether an adaptation of the instrument for routine colorimetry would be either satisfactory or practical.

The use of spectrum components is scarcely practical for color measurements, since instruments of sufficient accuracy employing spectrum components are very complicated and delicate. Reliable and precise control requires elaborate devices, and special precautions are required to prevent contamination of the spectrum components with stray light. Also, mixtures consisting of two or three spectrally pure components are so dissimilar to the continuous spectral distributions of most stimuli encountered in practice that serious discrepancies arise among the results obtained by different observers. These personal variations are caused by the varying color-mixture data of different individuals, the effect being greatest when stimuli of extremely different spectral compositions are compared. Although the color-mixture functions for different "normal" observers are frequently very different, these individual variations are rarely of practical importance. Visually similar stimuli of common occurrence are usually similar in spectral composition, and for them considerable variations of the color-mixture data cause negligible discrepancies in color matching. For this reason, average color-mixture functions can be used satisfactorily and standardized, even though considerable variations from the average curves are frequently encountered. This tolerance for personal variations of color vision is sacrificed when a colorimeter employing spectral components is adopted.

3. THREE-COLOR COLORIMETERS USING FILTERS

Trichromatic colorimeters employing spectrally selective filters are more suitable for

routine colorimetry than are instruments employing spectrum components. The light transmitted by such filters can be evaluated accurately and completely with no uncertainty corresponding to that arising from the stray-light contamination of spectrum components. The flux of filtered beams can be conveniently and accurately controlled and evaluated. Mixtures of filtered components are more nearly spectrally similar to ordinary color stimuli than are mixtures of spectrum components, and the disturbing effects of personal variations of color vision are therefore less serious than those encountered in the use of spectrum components. Colorimeters employing filters can be more simple, less delicate, and more portable than colorimeters employing spectrum components.

The first trichromatic colorimeter employing filters was developed by F. E. Ives.[807] The components were mixed by the use of a rotating crown of lenses, which caused the images of three filters to sweep successively across the pupil of the observer's eye. Regulation of the width of diaphragms placed over the filters provided for variation of the amounts of the three components. No provision was made for the measurement of colors which were too pure to be matched with mixtures of the filtered components.

Guild used a three-filter colorimeter[808] for the determination[615] of data confirming the results of Wright (and used as a basis for the 1931 I.C.I. standardization). Light enters Guild's colorimeter, Fig. 3,* through three sector-shaped windows, covered with red, green, and blue filters, respectively. The angular extent of the open sectors can be varied by rotatory movements of shutters. One end *C* of a periscopic prism rotates rapidly past the sectors. The light is reflected along the axis of rotation from the other end *D* of the prism. When the speed of rotation is high, the alternation of color of the light reflected along the axis cannot be perceived, and the resulting color corresponds to the mixture of the components. Light from the sample *S* under test illuminates the comparison side of the field of view *F*. Provision, *L–M–N–K–J*, is made for adding any of the filtered components to the test color, when required. The Guild colorimeter is useful for all kinds of color measurements, for routine specification and control, as well as for research on color vision. The results obtained with any trichromatic colorimeter can be converted into the I.C.I. system. Guild described the methods of calibrating instruments such as his own, and the conversion of the results into any system of colorimetric units. The I.C.I. specifications of the light transmitted by the three filters employed in the colorimeter can be computed from the spectrophotometric characteristics of the filters and of the light incident upon them. The readings secured with the colorimeter can then be expressed in terms of I.C.I. units by the use of the methods of computation of additive mixtures outlined in Chapter VII. In theory, the

[807] F. E. Ives, Trans. Illum. Eng. Soc. **3**, 627 (1908).
[808] J. Guild, Trans. Opt. Soc. Am. **27**, 106 (1925–26).

* Reprinted by permission of J. Guild.

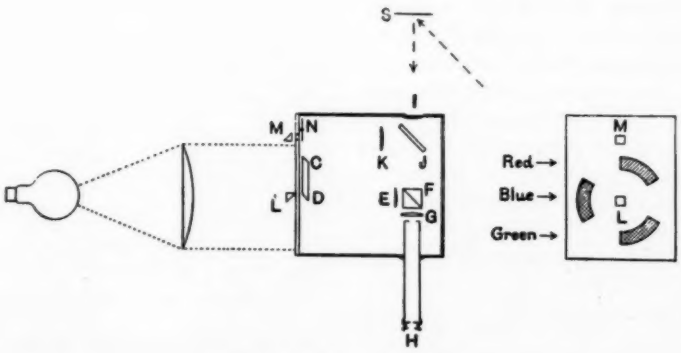


FIG. 3. Guild's colorimeter.

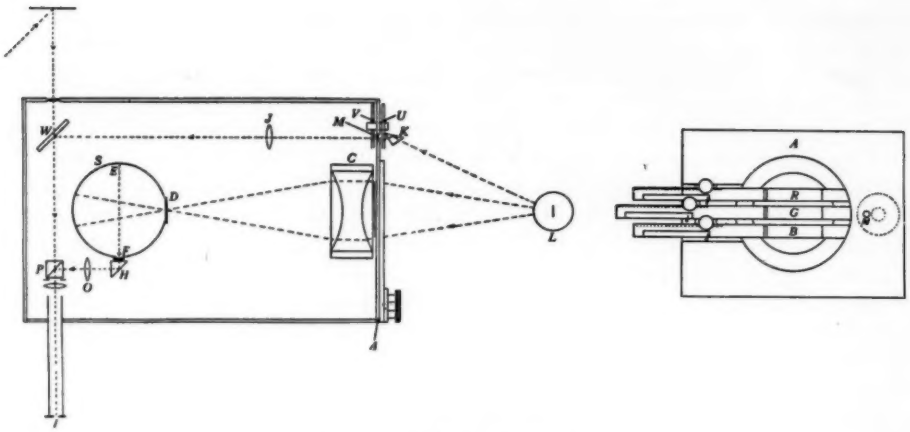


FIG. 4. Donaldson colorimeter.

results can be identical with the standard I.C.I. specifications only if the observer using the colorimeter has color-mixture functions indistinguishable from those of the standard I.C.I. observer. Such observers are rare, and the results obtained with any colorimeter are at best only good approximations to standard I.C.I. specifications.

Donaldson has designed a trichromatic colorimeter using filters, and free from moving parts.[809] The three filters *R*, *G*, *B* are located in three narrow rectangular openings in a diaphragm, Fig. 4,† placed in contact with a condenser lens *C*. The lengths of the rectangular apertures can be varied by adjusting the position of three sliding shutters. The light passing through the exposed areas of the three filters is mixed by diffusion in a hollow sphere *S* coated with magnesium oxide. An image of a portion of the interior of this sphere is formed in one-half of the field seen by the observer. The comparison half of the field may be illuminated with light from the source *L* which illuminates the sphere or with light from an external source. Transparent materials may be placed in the comparison beam for evaluation of the colors transmitted when the materials are illuminated with any desired quality of light. The external source of the comparison beam may consist of a reflecting surface, in which case the color evaluated is that of the light reflected by the surface. The Donaldson colorimeter is calibrated by the manufacturer in terms of the I.C.I. system.[810] The Donaldson colorimeter should be quite satisfactory and applicable to industrial use as well as to research in color vision. The accuracy of specifications secured by the use of this colorimeter is limited by the extent to which the color-mixture functions of the observer differ from the standard color-mixture functions.

4. COLORIMETERS FOR DIRECT DETERMINATION OF DOMINANT WAVE-LENGTH AND PURITY

Colorimeters designed for the direct determination of dominant wave-length and purity are usually of the "monochromatic-plus-white" type. P. G. Nutting designed an instrument of this kind, shown in Fig. 5.[811] The color under test is introduced into the photometric cube *C* by the beam *3*. This is matched by a mixture of spectrum light and a standard achromatic stimulus. The spectrum light is refracted from the collimator *1* by the prism *P*. The achromatic light is reflected by the prism *P* from the beam *2*. The wave-length of the spectrum light may be adjusted to the required value by adjustment of the angle of incidence of the beam from the collimator *1* on the prism *P*. The luminances of the spectrum light and the achromatic stimulus

[809] R. Donaldson, Proc. Phys. Soc. **47**, 1068 (1935).
† Reprinted by permission of R. Donaldson.

[810] Adam Hilger, Ltd., *The Donaldson Trichromatic Colorimeter*, Hilger Publication No. 250 (1936).
[811] P. G. Nutting, Bull. Bur. Stand. No. 9 (1913).

may be adjusted by varying the slit widths, by rotating sectors, or by rotating one of a pair of polarizing prisms placed just inside each slit. The ratio of the luminances of the constituents is determined by a separate photometric comparison. This step is the source of the greatest errors of colorimetry with this type of instrument, and this, together with the extreme sensitivity to observer differences, which is characteristic of all colorimeters employing spectrum components, is responsible for the virtual abandonment of the "monochromatic-plus-white" method of colorimetry. In Nutting's original instrument, the ratio of luminances was evaluated by direct heterochromatic comparison, and was subject to the errors and ambiguities inherent in that method of photometry. In a form of the instrument built by Adam Hilger, Ltd., a flicker photometer is provided for the photometric comparisons. Because of the narrow exit slit, and the use of pairs of Nicol prisms for flux control of the various beams, the luminance of the test field of this instrument is lower than the luminance of the test object. Consequently, the instrument is suitable only for the measurement of colors of fairly high luminance. Priest described an assembly of apparatus for measuring dominant wave-length and purity in which the luminance adjustments are accomplished by changing the distances of the sources from the comparison field.[812] The heterochromatic comparisons are made by use of a flicker photometer. The conditions of observation are more suitable for accurate determinations than in earlier instruments.

Guild has questioned the advisability of the introduction of a luminosity comparison in the determination of chromaticity.[813] Variations of the luminosity function may occur for various observers for whom the color-mixture functions are nearly standard. Color measurements by such observers using the "monochromatic-plus-white" method are likely to be quite incorrect, although normal specifications of chromaticity are found by such observers when using any method of colorimetry which does not involve photometric determinations.

[812] I. G. Priest, J. Opt. Soc. Am. **8**, 173 (1924).
[813] J. Guild, Trans. Opt. Soc. Am. **27**, 130 (1925–26).

This objection to heterochromatic photometry as a basis for colorimetry applies to an even greater extent to the direct photometric method of colorimetry.[814] In this method the tristimulus values are determined directly as ratios of luminances of the sample and a standard "white" surface when observed through three filters suitable for the conversion of the luminosity curve into each of the standard color-mixture functions in turn. For the determination of the value of Y, an observer with standard luminosity characteristics would use no filter, and the evaluation would be an ordinary heterochromatic luminance comparison, subject to all the errors of heterochromatic photometry. For the determination of the value of X, such an observer would require the use of a filter having

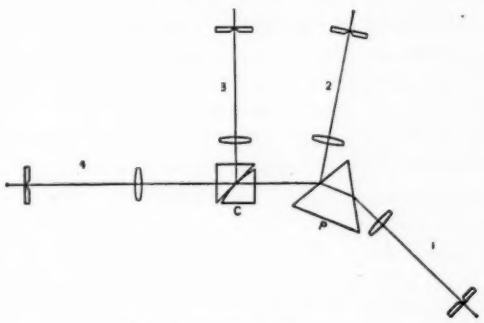


FIG. 5. Nutting colorimeter.

its spectral transmittance proportional to $\bar{x}/\bar{y}$, and for the determination of the value of Z a normal observer would require the use of a filter having its spectral transmittance proportional to $\bar{z}/\bar{y}$. Both of these filters would be exceedingly difficult to procure, and would have very low luminous transmittances. These filters would not be sufficiently selective to eliminate the chromaticity differences of various stimuli, and the difficulties of heterochromatic photometry would be increased rather than reduced by their use.

5. GUILD'S VECTOR COLORIMETER

To avoid heterochromatic photometry in colorimetry, and simultaneously to take advantage of the simplicity of two-component mixtures, Guild devised an instrument for the

[814] M. Richter, Das Licht **7**, 90 (1937).

direct determination of chromaticity by the establishment of two successive matches.[815] He named this instrument the vector colorimeter, because the chromaticity is determined by the intersection of two vectors in the chromaticity diagram, Fig. 6. Nutting's colorimeter, Fig. 5, could be employed as a vector colorimeter. For this purpose, a highly selective red filter would be introduced in the beam *2*. The location in the chromaticity diagram of the point *A*, representing the light transmitted by such a filter, can be determined by the use of the methods described in the previous chapter. A color match with any sample can be established by varying the wave-length of the spectrum component transmitted by the prism *P* and by adjusting the intensities in the beams *1* and *2*. The straight line *AB* passing through the points in the chromaticity diagram, Fig. 6, representing the two constituents of the mixture passes also through the point *S*, representing the chromaticity of the sample. No attention is paid to the luminances of the components necessary to establish the color match. The location of the point *S* on the line *AB* is the intersection with another line *CD* determined by the points *C* and *D* representing the components of another mixture also matching the sample. The components of this second mixture should be quite different from the components of the first mixture, and may be chosen in a systematic manner suggested by Guild. In all cases, one of the components, for instance that represented by *C*, may be produced by placing a filter in the beam *2* (or *3*) of the Nutting colorimeter or its equivalent.

Guild's vector colorimeter is well suited for the measurement of any stimulus having its energy concentrated in a single narrow region of the spectrum, but ordinarily negligible variations from standard color vision can cause serious errors in the measurement of stimuli of low purity, which are usually very dissimilar in spectral distribution to the mixtures synthesized in the vector colorimeter.

6. ROTATORY DISPERSION COLORIMETERS

Colorimeters have been described and constructed in which the colors produced by rotatory dispersion of polarized light have been used for the specification of chromaticity.[816] From the thickness of quartz plates and their arrangement between Nicol prisms, and from the orientation of the Nicol prisms, the energy distribution of the transmitted light from a given source can be calculated and the chromaticity of the resulting color can be determined. The calibration of such colors for a variety of plate thicknesses and orientations of the polarizing prisms was published by Buchwald,[520] who assumed the use of the I.C.I. standard Illuminant *B* and the standard observer. The calibration described by

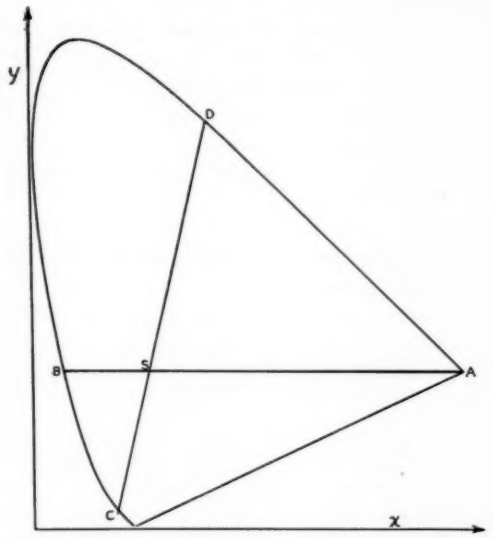


FIG. 6. Chromaticity diagram illustrating principle of vector colorimeter.

Buchwald indicates the general applicability of the method, but the data are not sufficiently complete to form the basis for a practical method of colorimetry. Since more than three independent variables are available in colorimeters based on the phenomenon of rotatory dispersion, an essential part of the design of a practical instrument would be a determination of the most favorable combination of variables for each region of the chromaticity diagram. The devising of a systematic method of procedure, and also the tabulation of graphical representation of the corresponding complete calibration for all chro-

[815] J. Guild, Trans. Opt. Soc. Am. **27**, 139 (1925–26).

[816] L. Arons, Ann. d. Physik **39**, 545 (1912).

maticities, would also be necessary. The advantages of the rotatory dispersion method of colorimetry are the absolute reproducibility of the instrument and the similarity to ordinary stimuli of the spectral distributions obtained. The most serious objections to this type of colorimeter arise from the very rapid and erratic variations of luminance which frequently accompany moderate changes of chromaticity. Doubtless, the most objectionable of these variations could be avoided by the use of an advantageously designed apparatus and system of operation, but such a method has not yet been made available for general colorimetry. Priest's rotatory dispersion colorimeter[819, 817] for the measurement of color temperature has accomplished the objectives outlined, but for only the very limited class of colors which are similar to the colors of light from complete radiators at various temperatures.

7. EMPIRICAL COLORIMETERS

Empirical colorimeters in which light from a source is passed in succession through variable combinations of selectively absorbing materials have been used extensively for color specification. These colorimeters occupy the borderline between colorimeters designed for the specification of color and those designed to use color as an indication of chemical constitution. Many of these colorimeters, frequently called "subtractive colorimeters," have been used for both purposes. In many cases the distinction is difficult to maintain because these instruments "specify" colors only in terms of their own empirical and arbitrary scales, and the specifications are frequently used in place of more fundamental specifications of chemical composition and physical state.

The Lovibond tintometer is one of the oldest established empirical colorimeters.[818] In it the light from the sample is compared with light transmitted through a combination of selectively absorbing glass filters. The filters of the combination are selected by trial from sets of graded filters until the best obtainable color match is established between the color of the light from the sample and the color of the light transmitted by the combination of filters. The glasses are in three series: One series (called yellow) absorbs principally blue light, a second series (called red) absorbs principally green light, and a third series (called blue) absorbs principally orange light. The glasses in each series are numbered approximately in proportion to their densities, so that a combination of any two members of a series is approximately equivalent to a single glass with a scale number equal to the sum of the scale numbers of the other two. The color of any sample is specified by the totals of the numbers of the glasses of each of the series necessary to produce a color match. Since the scale numbers are not strictly additive,[819] combinations consisting of the fewest glasses are preferable. Spectrophotometric studies of the glasses have been made at several laboratories, notably at the National Bureau of Standards.[820] These studies indicate that the densities of the glasses supplied in any one set are not accurately proportional to the scale numbers engraved upon them,[821] and that glasses from different sets differ considerably although engraved with the same nominal values.[822] Consequently, the results secured by the use of different sets of tintometer glasses are not accurately comparable, unless the glasses are recalibrated on the basis of spectrophotometric measurements. The readings obtained with any set of tintometer glasses can be converted into I.C.I. units only as a result of complete spectrophotometric analyses of the glasses. Trichromatic coordinates have been computed from such spectrophotometric measurements of a representative set of tintometer glasses.[823, 824] Illuminant *B* was assumed for these computations, and a similar set of computations for Illuminant *C* is in preparation. The general

[817] I. G. Priest, J. Opt. Soc. Am. **4**, 448 (1920); J. Opt. Soc. Am. **5**, 178 (1921).

[818] J. W. Lovibond, J. Soc. Dyers and Colourists **3**, 186 (1887); *Measurement of Light and Colour Sensations* (George Gill and Sons, London, 1893).

[819] G. W. Haupt, Oil and Soap **15**, 282 (1938).

[820] K. S. Gibson, F. K. Harris, and I. G. Priest, Sci. Pap. Bur. Stand. No. 547, **22**, 1 (1927).

[821] K. S. Gibson and G. W. Haupt, J. Research Nat. Bur. Stand. **13**, 433 (1934).

[822] I. G. Priest, D. B. Judd, K. S. Gibson, and G. K. Walker, Bur. Stand. J. Research **2**, 793 (1929); G. K. Walker, J. Research Nat. Bur. Stand. **12**, 269 (1934).

[823] The Tintometer (Ltd.), *Colorimetry* (Milford, Salisbury, England, 1939 and 1941).

[824] R. K. Schofield, J. Sci. Inst. **16**, 74 (1939).

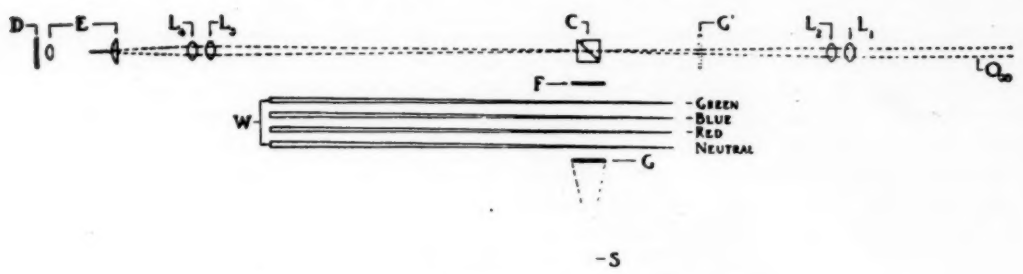


FIG. 7. Eastman subtractive colorimeter.

use of such typical values will introduce discrepancies corresponding to the differences between the particular glasses used and the glasses from which the spectrophotometric data and colorimetric computations were determined.

Each set of Lovibond glasses can be considered to be a set of color standards. Since different sets are not matched within limits negligible in comparison with the nominal differences between successive glasses, the Lovibond system cannot be considered to provide a system of universally equivalent standards. It seems more appropriate, therefore, to discuss the Lovibond system as a subtractive colorimeter, rather than as a system of color standards. The optical system of the tintometer is exceedingly simple, but the considerable separation of the comparison fields is unfavorable for accurate color comparisons. The Lovibond glasses are usually employed in other more suitable optical systems.[801, 822]

The tintometer has proved useful for the grading of specific classes of materials, the colors of which vary only within a limited range, for which sets of a few glasses can be specially selected and standardized.[822] In its most usual applications, therefore, the Lovibond system is more closely related to colorimeters which employ color as an indication of composition and quality, and only by a considerable relaxation of precise terminology can it be said to measure color.

The Eastman subtractive colorimeter,[825] represented in Fig. 7, is similar in principle to the Lovibond system, but employs wedges *W* of three selectively absorbing materials. The density of each of the absorbers increases continuously from one end of the wedges to the other. One of these wedges (minus green) absorbs principally green light, the second (minus blue) absorbs principally blue light, and the third (minus red) absorbs principally red light. A fourth wedge (neutral) is used to secure a luminance match. The colors of most samples *O* can be matched by sliding the wedges until different thicknesses of the three absorbing media are in the path *S–C* of the comparison light. The scales attached to the wedges are calibrated in terms of distance along the wedges. The Eastman colorimeter is no longer available from the manufacturer. The readings from this type of colorimeter could be converted into I.C.I. values if the spectrophotometric curves for a number of positions on each wedge were determined. Such data are not available for the Eastman colorimeter.

A subtractive colorimeter designed for very precise measurements of small chromaticity differences was constructed and described by D. B. Judd.[826] This colorimeter employs two double wedges, one of greenish and the other of yellowish glass. These wedges are placed in the comparison beam of a photometer in which two surfaces, viewed in juxtaposition in a Lummer-Brodhun cube, are illuminated by a single lamp movable to any position between the surfaces. The differences in wedge settings can be calibrated in terms of differences of coordinates in any chromaticity diagram, provided the spectral reflectances of the comparison surface are known approximately and the spectral transmissivities of the wedges are known accurately. Because of the large patterned field and high luminance, this instrument takes full advantage of the

[825] L. A. Jones, J. Opt. Soc. Am. **4**, 420 (1920).

[826] D. B. Judd, Textile Research **9**, 253 (1939).

chromaticity sensibility of the observer. It is not applicable to highly selective samples because the nearly homogeneous energy reflected by such samples is too little changed in spectral distribution by passage through the wedges.

The Eastman color densitometer[827] is essentially a modification of the Eastman subtractive colorimeter. The color densitometer was designed specifically for a very specialized problem, the measurement of the effective concentrations of three selectively absorbing materials used in color photography. This is accomplished by making the wedges of the instrument with the same absorbing materials that are used in the photographic process. The light passing through any part of a color photograph is passed also through suitably selected parts of not more than two of these three wedges, and through a neutral wedge. The densities of the wedges necessary to make the light reaching the eye of the observer match that from a neutral comparison field constitute a specification of the color in the photograph, in terms of the constituent colorants of the photographic image. Insofar as the instrument is intended to be used solely for the evaluation of the concentrations of the image components, it should be classified with chemical colorimeters, and as such it need not be discussed here. It is, however, equivalent in principle to the other subtractive colorimeters already discussed, and with appropriate rearrangement it would be equally applicable to general colorimetry. The results expressed by its specialized scales could be converted into I.C.I. specifications by the use of calibrations similar to those discussed in the cases of the Lovibond and Eastman subtractive colorimeters.

8. DISK COLORIMETERS USING COLOR STANDARDS

Colorimeters have been designed for use with several systems of color standards. The systems of color standards will be discussed in another section of this chapter. Disk colorimeters have been designed for use with the Munsell system of color standards.[828] In these instruments the well-known method of mixing colors by rotating sectors has been modified in the interest of simple construction and rapid operation. The sectors of paper reflecting light of different colors are placed on a stationary disk, so that their proportions can be changed continuously, and a rapidly rotating prism scans an annular area of the sector disk. The results are expressed as interpolations between the Munsell notations of the standard Munsell papers used for the sectors.[829] Spectrophotometric analyses of all of the Munsell standards are available. Specifications based on these spectrophotometric curves, in terms of the I.C.I. observer and standard Illuminant *C*, have been published.[830–832] Consequently, readings obtained with disk colorimeters, employing standard Munsell papers and the equivalent of Illuminant *C*, are convertible into I.C.I. units.[833] The effects of personal variations from the standard color-mixture functions are usually minimized in disk colorimetry by employing for each problem sectors which are spectrophotometrically very similar to the samples being matched. Revisions of the notations of the Munsell standards[834] do not necessitate any revisions of the I.C.I. specifications derived from the use of Munsell standards in disk colorimetry. In the derivation of I.C.I. specifications from disk mixtures, the notations of the sectors, whether in terms of the present Munsell system or any revision of that system, serve simply as means of identification of the standards employed in the sectors. Any notation which unambiguously designates the materials used for the sectors will also serve to identify the corresponding spectrophotometric data, from which the I.C.I. specifications of the disk mixture can be computed.

Disk colorimetry need not be limited to the use of Munsell papers; any material that can be cut

[827] R. M. Evans, J. Soc. Mot. Pict. Eng. **31**, 192 (1938).
[828] D. Nickerson, J. Opt. Soc. Am. **21**, 640 (1931). J. W. Forrest, Paper Trade J. TAPPI Section **97**, 266 (Nov. 30, 1933).
[829] D. Nickerson, J. Opt. Soc. Am. **21**, 643 (1931); J. Opt. Soc. Am. **25**, 253 (1935).
[830] J. J. Glenn and J. T. Killian, J. Opt. Soc. Am. **30**, 609 (1940).
[831] K. L. Kelly, K. S. Gibson, and D. Nickerson, J. Opt. Soc. Am. **33**, 355 (1943).
[832] W. Granville, D. Nickerson, and C. E. Foss, J. Opt. Soc. Am. **33**, 376 (1943).
[833] D. Nickerson, U. S. Dept. Agriculture, Agricultural Marketing Service, Washington (1938).
[834] S. M. Newhall, J. Opt. Soc. Am. **30**, 617 (1940); S. M. Newhall, D. Nickerson, and D. B. Judd, J. Opt. Soc. Am. **33**, 385 (1943).

into suitable disks, and for which I.C.I. tristimulus values are available, can be used as standards. The instrument is an additive colorimeter, and colorimetric results are obtained directly in terms of I.C.I. tristimulus values.

9. PHOTOELECTRIC COLORIMETERS

Instruments called "photoelectric colorimeters" have been described frequently in recent years. The feature of a photoelectric colorimeter which attracts most favorable attention is the independence and separability of simultaneous photo-cell responses. In this respect, the photocell appears to many workers to be a more convenient receptor for colorimetry than the human eye. Separable responses of the eye have never been demonstrated, although the existence of separate responses must be assumed in order to account for the observable facts of color vision.[835] Guild[202] and Gibson[836] have discussed the conditions which must be fulfilled by any photoelectric instrument before it can be properly called a colorimeter. No instrument which has receptor characteristics essentially different from visual color-mixture functions can be related to the essential properties of the eye, any more than could a system of photometry which took no account of the luminosity curve.

The conditions to be fulfilled by any photoelectric instrument worthy of the name colorimeter[836] are that the products of the spectral sensitivity of each of three photo-cells times the spectral transmittance of any filter used with it should be accurately proportional throughout the spectrum to some linear combination of the standard I.C.I. color-mixture curves. A single photo-cell can be substituted for the three if successive responses through three (or more) filters are recorded and if the sample color remains unchanged during the time required for the three measurements. In general, the responses of these colorimeters may be converted to I.C.I. specifications by the calculation of corresponding linear combinations of the photocell responses. Care must be taken to insure that the indications of photoelectric colorimeters are strictly proportional to the flux to be measured.[837] No instrument exactly fulfilling all of the necessary conditions has ever been built,[565, 838, 839] although several instruments have been described which are reported to give good approximations to the desired accuracy.[840–843] The details of construction, the accuracy claimed, and the limitations of use of these instruments are discussed in the original papers and in published surveys.[565, 838] Because of the manifold difficulties attending the fulfillment of all the necessary conditions, all instruments of this kind should be subjected to very careful and critical scrutiny before adoption. Very little compromise with the fundamental requirements can be tolerated.[836]

10. TRANSPARENT COLOR STANDARDS

Material standards of color are convenient and references to such standards are often preferred to specifications based on the I.C.I. standardizations.[801, 844] Tangible color standards are readily comprehended by those who have no acquaintance with colorimetry. Material samples may serve as secondary standards even when the primary standards are specified in terms of the I.C.I. system.[845] Specifications of color in terms of any system of material standards consist of the identification of the particular member of the system which matches the color being specified. The arrangements of the various standards of any complete set in related series provide systematic means of identification, in terms of which the specifications of any color may be expressed. A great number of systems of color standards have been devised.[846–848] Only widely used systems, which have color gamuts suffi-

[835] J. Guild, J. Sci. Inst. **11**, 69 (1934).
[836] K. S. Gibson, Instruments **9**, 309, 335 (1936); Nat. Bur. Stand. Letter Circular LC 545 (March 8, 1939).
[837] R. S. Hunter, J. Opt. Soc. Am. **30**, 536 (1940).
[838] Institute of Paper Chemistry, Paper Trade J., TAPPI Section **105**, 285, 293, 356, 380 (1937); **107**, 231, 275 (1938).
[839] J. A. VandenAkker, J. Opt. Soc. Am. **27**, 401 (1937).
[840] J. W. Perry, J. Sci. Inst. **15**, 270 (1938).
[841] R. S. Hunter, J. Opt. Soc. Am. **28**, 51, 179 (1938); **31**, 463 (1941).
[842] B. T. Barnes, J. Opt. Soc. Am. **29**, 448 (1939); J. Opt. Soc. Am. **31**, 463 (1941).
[843] G. T. Winch and E. H. Palmer, Trans. Illum. Eng. Soc. (London) **2**, 137 (1937).
[844] D. B. Judd, Trans. Illum. Eng. Soc. (New York) **36**, 336 (1941).
[845] D. B. Judd, Am. J. Psychol. **52**, 418 (1939).
[846] A. Ames, Jr., J. Opt. Soc. Am. **5**, 160 (1921).
[847] K. S. Gibson, J. Soc. Mot. Pict. Eng. **28**, 388 (1937).
[848] F. L. Dimmick, Psychol. Bull. **35**, 473 (1938).

ciently extensive to represent almost all of the colors exhibited by non-self-luminous objects can be discussed here.

The Lovibond system has been discussed in connection with subtractive colorimeters.[818–824] References [804] have been given to the literature of the system based on the Arny solutions, which are used almost exclusively for the grading of materials and for the specification of the colors exhibited by indicators used in chemical analysis.[849] The Arny solutions contain selectively absorbing materials, the concentrations of which are adjusted to secure a color matching the color of the sample. The color of the sample is then specified in terms of the concentrations of the standard solutions necessary to produce the color match. A solution of cobalt chloride serves to absorb principally green components of the light transmitted. An aqueous solution of ferric chloride absorbs blue light almost exclusively, and a solution of copper sulfate in one percent hydrochloric acid absorbs principally in the red region of the spectrum. All colors except certain reds and blues of high purity can be produced with this group of three solutions. Various supplementary solutions are used for special problems.[804]

11. COLOR STANDARDS OF REFLECTING SURFACES

Sets of material standards usually consist of pigmented or dyed surfaces. The various systems differ principally in the arrangements of the standards in series and groups of series which facilitate identification and specification of the standard matching any sample.

12. MUNSELL SYSTEM

In the Munsell system, the color standards are grouped in three different ways.[850, 851] All samples ordinarily perceived in daylight as having approximately the same hue are grouped together. Twenty hues are exhibited by the samples in the 1929 edition of the *Munsell Book of Color*.[852] An additional series, consisting of gray samples, can be combined equally well with each of the groups of constant hue. The gray samples furnish terminal members for several series into which each group of constant hue is subdivided. Each of these series contains standards ordinarily perceived in daylight as having approximately equal lightness (Munsell value). The standards in each of these series are arranged in accordance with the degree of apparent difference from the member of the gray series perceived as having the same lightness. The characteristic of the color perception according to which the standards are arranged in each series of constant lightness is saturation (Munsell chroma).

The standards of the Munsell system have been selected so that approximately equal saturation differences are ordinarily perceived in daylight between successive members of each of these series. As a consequence, each group of standards of the same hue can also be arranged in several series consisting of standards perceived as having approximately equal saturation (chroma). The standards in each series are arranged according to lightness. The members of the gray series, and of the corresponding series of equal lightness, have been selected so that approximately equal lightness differences are perceived between successive members of each of the series of constant saturation. Consequently, these two types of arrangements of standards of any one hue can be combined and represented simultaneously in a rectangular array. The gray scale is arranged in a vertical series with white at the top and black at the bottom. Each series of constant lightness is arranged horizontally, the various standards in the series being arranged in order of saturation, extending from the corresponding member of the gray series to the most saturated standard of the correct hue and lightness. In such a rectangular array, all standards ordinarily perceived in daylight as having the same saturation fall into a vertical series, arranged according to lightness in the same manner as the gray scale which is parallel to it. The arrays for complementary hues are frequently placed together, with the scales of increasing

[849] United States Pharmacopoeia, articles on cod-liver oil and carbonization tests with sulphuric acid.

[850] A. H. Munsell, *A Color Notation* (Ellis, Boston, 1905, 1907, 1913, 1916); (Munsell Color Company, Baltimore, 1923–1926, 1936).

[851] D. Nickerson, J. Opt. Soc. Am. 30, 575 (1940).

[852] Munsell Color Company, *Munsell Book of Color* (Baltimore, 1929).

saturation extending in opposite directions from a gray scale which is shared by the arrays for the complementary hues.

The arrays for various hues are related in a manner that is represented by a three-dimen-

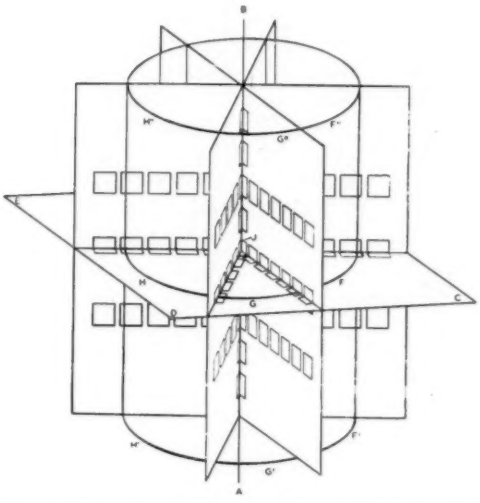


FIG. 8. Three-dimensional arrangement of Munsell system.

sional arrangement, indicated in Fig. 8, by planes intersecting in a single line *AB*. The common intersection of the planes representing the different hues represents the gray series. The standards of the Munsell system have been selected so that equal hue differences are perceived between successive members of a series of standards having the same Munsell value and chroma. This equal spacing of the perceived hues of the groups of Munsell standards is represented by orienting the planes of the corresponding arrays so that equal angles are formed by the intersecting planes representing successive hues.

The arrangement of the standards in the manner described places them at the corners of the unit cells of a cylindrical coordinate system. All of the standards having any specified lightness (value) are represented on a plane, e.g., *CDE*, perpendicular to the axis. The planes representing different lightnesses are arranged in the order of lightness, corresponding to the arrangement of the standards, e.g., *J*, in the gray series *AB*. In any one of these planes, standards of constant hue are arranged along radial lines (as represented in Fig. 8) with complementary hues on diametrically opposed radii. Standards of different hue but equal Munsell value and chroma are represented on circles (e.g., *FGH* concentric with the point *J*, representing the gray). The planes passing through the axis, the planes perpendicular to the axis, and the cylindrical surfaces (e.g., *G'H'F'–G''H''F''*) concentric with the axis represent standards ordinarily perceived in daylight as having approximately constant hue, lightness, and saturation, respectively. The equal separations of the points representing Munsell standards arranged in series according to any one of these attributes correspond to the approximately equal differences which are perceived between the successive color standards. The form of the coordinate system is therefore fairly representative of the perceptual relations between the color standards of the Munsell system viewed in daylight against a light gray or white background. The surface enclosing all of the points representing standards included in the present Munsell system is very irregular. This irregu-

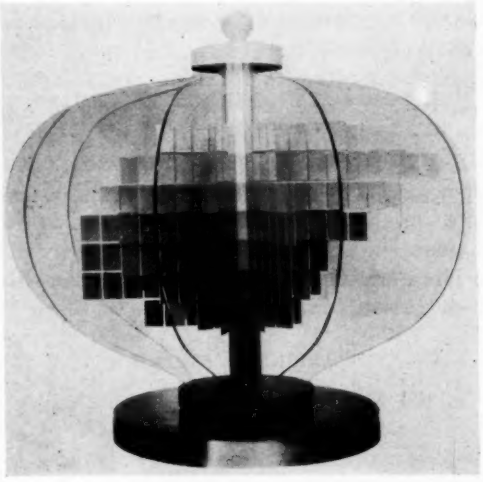


FIG. 9. Munsell color tree constructed by Arthur S. Allen.

larity represents the limitations of the pigments used, and is not characteristic of color vision. The coordinate system defined by the Munsell standards is unlimited and colors produced by new pigments, yielding saturations and light-

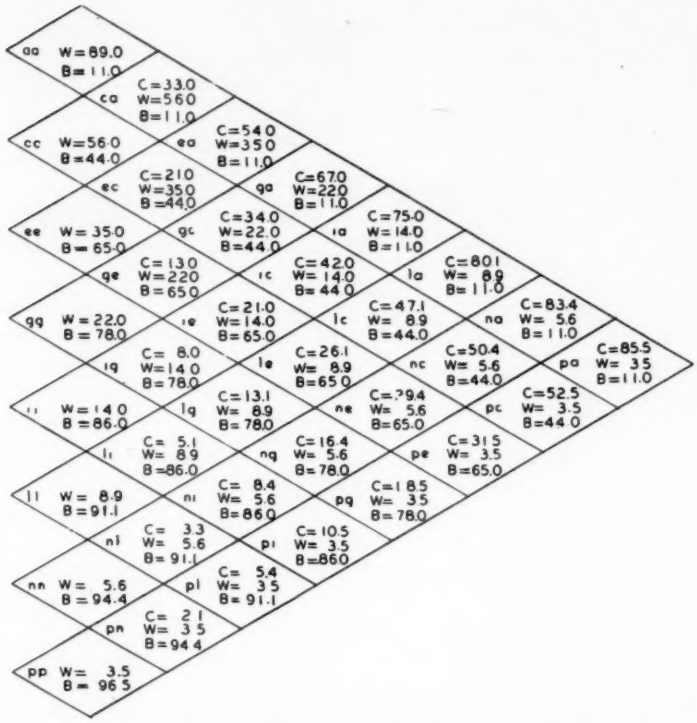


FIG. 10. Triangular arrangement of samples in Ostwald system.

nesses greater than those now attainable, can be accommodated in the system.

The arrangement of the series of constant hue and lightness, with varying chromas, along lines perpendicular to the gray axis, Fig. 9, suggests the origin of the common description of the Munsell system as a "color tree."

The 1942 edition of the Munsell charts has doubled the number of available hues (and samples). The standards in the *Munsell Book of Color* were selected on the basis of extensive investigations[736, 757, 853–859] of the perceptual differences in the various series. The arrangements of the standards in the *Munsell Book of Color* have been reviewed in comparison with more recent measurements and observations, by a subcommittee of the colorimetry committee of the Optical Society of America.[834]

Spectrophotometric curves for the standards of the 1929 *Munsell Book of Color* were prepared in 1934 in the color measurements laboratory of the Massachusetts Institute of Technology. Colorimetric specifications were computed from those curves by Glenn and Killian,[830] using the standard I.C.I. color-mixture data and standard Illuminant *C*. Similar data were published by Kelly, Gibson, and Nickerson for the standard Illuminants *A*, *B*, and *C*.[831] The spectrophotometric data and I.C.I. specifications of the samples of the twenty intermediate hues of the 1942 edition, and all supplementary Munsell samples, were determined by Granville, Nickerson, and Foss.[832] A reliable basis is therefore available for the transformation of color specifications expressed in the Munsell notation into I.C.I. specifications.

[853] A. E. O. Munsell, L. L. Sloan, and I. H. Godlove, J. Opt. Soc. Am. **23**, 394 (1933).
[854] I. H. Godlove, J. Opt. Soc. Am. **23**, 419 (1933).
[855] I. G. Priest and F. G. Brickwedde, J. Opt. Soc. Am. **13**, 306 (1926); **28**, 133 (1938).
[856] I. G. Priest and D. B. Judd, J. Opt. Soc. Am. **14**, 137 (1927).
[857] D. B. Judd, J. Opt. Soc. Am. **22**, 72 (1932).
[858] D. B. Judd, J. Opt. Soc. Am. **23**, 35 (1933).
[859] K. S. Gibson and D. Nickerson, J. Opt. Soc. Am. **30**, 591 (1940).

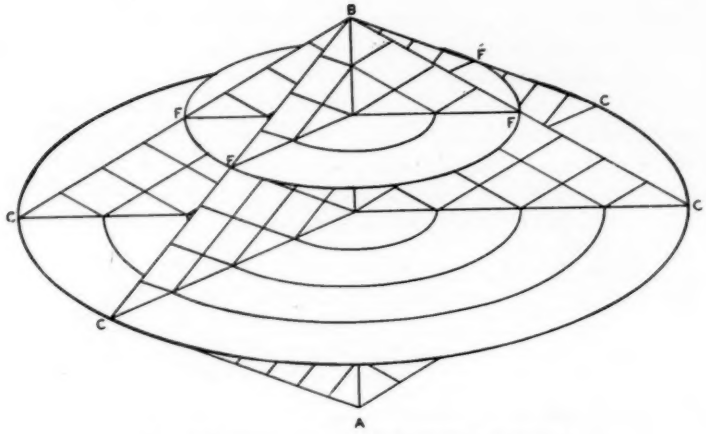


FIG. 11. Schematic diagram of Ostwald color solid.

Nickerson and Granville[860] have reported on the relation of the psychological concepts of hue and saturation (based on measurements of the Munsell colors) to the psychophysical concepts of dominant wave-length and purity. Tyler and Hardy,[861] and Gibson and Nickerson[859] investigated the consequences of the rules of additive mixture which Munsell originally postulated for his system. In terms of a system derived from those rules, Tyler and Hardy presented data for this system and determined for every value of each hue the maximum chromas which cannot be exceeded by any possible pigmented or dyed surfaces. These limits define for the early form of the Munsell system the color solid which has long intrigued color theorists.[862] It should be noted that the papers by Tyler and Hardy, and by Gibson and Nickerson refer to the early form of the Munsell system, not to the revised system which is illustrated in the 1929 *Munsell Book of Color*.

13. OSTWALD SYSTEM

The Ostwald system[863–869] of color standards is represented by twenty-four charts,[868] each of which contains twenty-eight standards of approximately constant dominant wave-length, but differing in purity and luminous reflectance. The samples are arranged on each chart in a triangular array (as in Fig. 10). A series of gray samples, varying from nearly white (*aa*) at the top to nearly black (*pp*) at the bottom, is arranged along a vertical side of each triangle. On each chart, the sample having the maximum purity is placed at the third vertex (*pa*) of the triangular array.

A series of samples (*pa*, *na*, *la*, *ia*, *ga*, *ea*, *ca*) of decreasing purity and increasing luminous reflectance is arranged at the positions indicated along the upper side of the triangular array. This series is called the scale of tints of the purest color (*pa*). A series of samples (*pc*, *pe*, *pg*, *pi*, *pl*, *pn*) of decreasing luminous reflectance and purity is arranged along the lower side of the triangle. This series is called the scale of tones of the purest member of the series (*pa*). Other series (e.g., *pe*, *ne*, *le*, *ie*, *ge*, *ee*) containing fewer

[860] D. Nickerson and W. Granville, J. Opt. Soc. Am. **30**, 159 (1940).
[861] J. E. Tyler and A. C. Hardy, J. Opt. Soc. Am. **30**, 90, 587 (1940).
[862] E. Schrödinger, Ann. d. Physik **62**, 603 (1920).
[863] W. Ostwald, *Die Farbenfibel* (Unesma, Leipzig, 1916); Zeits. f. physik. Chemie **91**, 129, (1916); Physik. Zeits. **17**, 322, 352 (1916); *Die Farbenlehre, I-V*, (Unesma, Leipzig, 1917, 1919, 1922); *Die Farbschule* (Unesma, Leipzig, 1919).
[864] W. Ostwald, *Color Science*, I and II, authorized translation of *Die Farbenlehre*, with an introduction and notes by J. Scott Taylor (Winsor and Newton, London, 1931–1933).
[865] W. Ostwald, *Die Grundlagen der messenden Farbenlehre* (Barth, Leipzig, 1921); also Zeits. f. tech. Physik **1**, 173, 261 (1920); **2**, 145 (1921).
[866] W. Ostwald, *Der Farbatlas* (Leipzig, 1917).
[867] W. Ostwald, *Der Farbkörper* (Leipzig, 1919, 1926); *Farbkunde* (S. Hirzel, Leipzig, 1923).
[868] W. Ostwald, *Der Farbnormenatlas* (Unesma, Leipzig, 1920, 1923, 1925).
[869] J. S. Taylor, *The Ostwald Color Album, A Complete Collection of Colour Standards for Use in Colour Specification and the Study of Colour Harmony* (Winsor and Newton, London, 1931).

samples are arranged parallel to the upper side of the triangle. Such a series of tints of the same tone (*pe*) is called an isotone series. Other series (such as *la*, *lc*, *le*, *lg*, *li*, *ll*) are arranged parallel to the lower side of the triangle. Such a series of tones of the same tint (*la*) is called an isotint series. The members of all of these series are chosen so that all the samples which fall in a vertical series (e.g., *ga*, *ic*, *le*, *ng*, *pi*) parallel to the gray series have equal purities, varying only in reflectance. Such a series is called an isochrome, or shadow series, and all members of the series are called shades of the same color.

The standards of the Ostwald system are specified in terms of quantities called "white content" W, "black content" B, and "color content" C. These quantities correspond in a general way to the fractions of the area of a spinning disk covered with white, black, and colored sectors, and producing a visual match for each standard. Improved techniques for measurement were devised and employed by Ostwald, but the purposes of these were to realize the results which can be conceived in terms of ideal disk mixtures. For example, a perfect black sector cannot be secured, and Ostwald employed a polarization method to secure the desired effect. Similarly, the highest purities obtained with pigmented surfaces can be increased by the introduction of more efficient pigments and do not in general correspond to any fundamental limits. Ostwald devoted much thought to the discovery of the fundamental limits of the purity of reflecting surfaces. His most important contribution to the science of colorimetry was his discovery that the colored surfaces which would have the greatest effectiveness (moment, or tinting strength) in disk mixtures would:

(1) reflect completely between complementary wave-lengths of the spectrum and absorb completely all other wave-lengths, or,

(2) absorb completely between complementary wave-lengths and reflect completely all other wave-lengths, or,

(3) reflect completely all wave-lengths greater than some wave-length which has no spectrum complementary and completely absorb all other wave-lengths, or,

(4) reflect completely all wave-lengths less than some wave-length which has no spectrum complementary and reflect perfectly all other wave-lengths.

Surfaces of this kind, which Ostwald called semichromes, and which he defined as the full colors ($C=1.0$) of his system, are unrealizable in practice, but they can be made the basis for scales of color content according to Ostwald's definitions. Ostwald attempted to base his measurements on such ideal scales. The tremendous variation of luminous reflectance (0.06 to 0.94) among the Ostwald semichromes was responsible for the invention of a concept of natural black content of the cold full colors, in particular the violets, blues, and bluish greens. Although Ostwald gave estimates of these natural black contents, he did not specify any method of measuring those quantities, and the natural black contents are not included or otherwise indicated by the black contents assigned to the standard colors in the charts of the Ostwald system. Discussions of this and other concepts proposed but not clarified by Ostwald have been frequent.[870–878] Many of the more specific interpretations of Ostwald's ideas have not gained acceptance among orthodox adherents of the Ostwald system, and frequent disputes and misunderstandings have arisen. Consequently, great difficulty is experienced in interpreting specifications of color given in terms of the Ostwald system.

In practice, the scales of black, white, and color content are defined by the values of these quantities assigned to the colored samples contained in various editions of the charts. The first Ostwald atlas[866] contained over 2500 samples, 100 hues, with 16 gray steps for 24 hues and eight steps for the remaining hues. Later editions[868–870] include only 24 hues, with eight steps in the gray scale and the scales of tints and tones

[870] *Color Harmony Manual* (Container Corporation of America, Chicago, 1942).
[871] K. W. F. Kohlrausch, Physik. Zeits. **21**, 473 (1920).
[872] C. Schaefer, Physik. Zeits. **27**, 347 (1926).
[873] R. Luther, Zeits. f. tech. Physik **8**, 540 (1927).
[874] A. Klughardt, Zeits. f. tech. Physik **8**, 299 (1927); **9**, 382 (1928); **10**, 101 (1929).
[875] C. Schaefer and H. Pese, Physik. Zeits. **31**, 720 (1930).
[876] K. Miescher, Zeits. f. tech. Physik **11**, 66 (1930); **11**, 233 (1930).
[877] C. Schaefer and H. Pese, Physik. Zeits. **32**, 1 (1931).
[878] M. Richter, Zeits. f. tech. Physik **12**, 582 (1931).

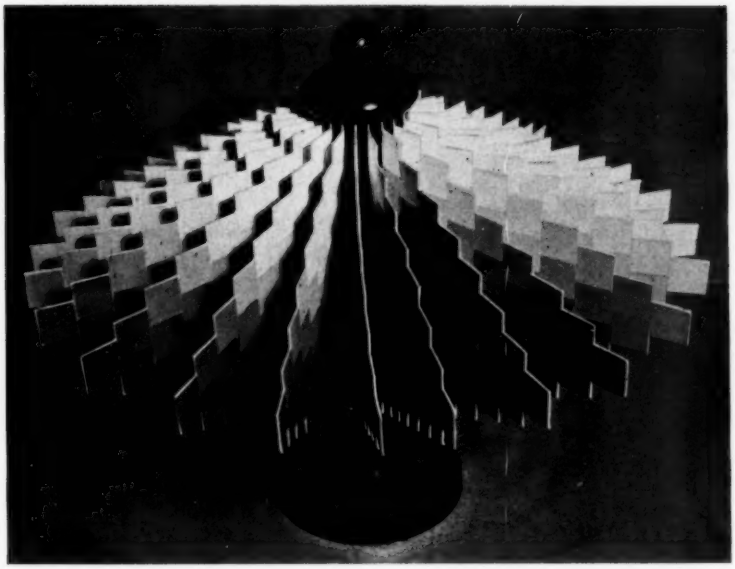


FIG. 11a. Model of Ostwald color solid constructed by E. G. Jacobson.

of the purest sample. The values of the black, white, and color contents assigned to the standards located at various positions in the Ostwald charts are shown in Fig. 10. The black content B is the same for all members of any isotone series, arranged parallel to the upper side of the triangle. The white content W is the same for all members of any isotint series, placed parallel to the lower side of the triangle. The ratio of color content C to white content W is constant for all members of any shadow series, placed in a vertical line, parallel to the gray series. In principle, and to a close approximation in the published charts, the purity is the same for all members of any shadow series.

The members of the gray series (*aa* to *pp*) are assigned white contents decreasing in constant proportion, and the reflectances of the standards of this series correspond quite accurately to this geometric progression. Ostwald chose this rule for the samples of the gray scale on the basis of Fechner's law for brightness discrimination, and because the resulting series of grays exhibited steps more nearly equal subjectively than a series having constant differences of reflectance. The color charts based on Fechner's law are called logarithmic charts. Each color standard is assigned the same white content as the member of the gray series included in the same isotint series, and the same black content as the member of the gray series falling in the same isotone series. The color content assigned is sufficient to make the total unity in all cases, $W+B+C=1.0$. Ostwald attributed great significance to this fact, although it had a simple and obvious origin in the disk mixtures upon which his conceptions were based. The three corresponding sectors completely covered the disk, and consequently the sum of the fractions of the area of the disk occupied by these sectors was always unity.

The variations in any isotone series correspond to increases in the white content at the expense of color content alone. Variations in any isotint series correspond to increases in the black content at the expense of color content alone. Variations in any shadow series correspond to increases in the black content at the simultaneous and proportionate expense of both white and color contents.

Although the Ostwald standards were selected on the basis of extensive measurements, they do not provide very accurate representations of the scales defined by Ostwald. Samples bearing the same notations in the various editions of the

charts exhibit considerable differences. These discrepancies appear to arise from endeavors to eliminate irregularities in the scales defined by the earlier charts, and from the use of improved pigments, of greater permanence and yielding colors of greater purity. Various editions show the samples either on triangular charts of the kind just described, or on circular charts showing the hue circle for a given white-black-color content. No edition presents the charts in more than one way. When the samples of one edition are arranged on a different diagram, the discrepancies are clearly revealed. The most recent edition, in which the standards are made of durable modern materials, is the most satisfactory according to this test for self-consistency.[870]

Kohlrausch[879] measured sixty Ostwald standards and computed tristimulus specifications for them, using the color-mixture data published by König and Dieterici.[880] Rösch[881] determined the spectrophotometric curves for 109 of the standards contained in the 1925 *Farbnormenatlas*[868] and published the trichromatic data for these based on the modification of König's color-mixture curves proposed by Ives.[882]

Ostwald's color solid is constructed by placing the logarithmic charts in the order of the hues of the standards, as indicated in Figs. 11 and 11a. The gray scales are all placed together to form the axis of a double cone,[883–885] of which each pair of charts for complementary hues represents a meridional plane. The upper sides *CB* of the triangles, containing all the tints of the purest standards (*pa*) fall on a conical surface with the white standard at the vertex *B*. The lower sides *AC* of the triangles, containing all the tones of the purest standards, are arranged on another conical surface, the vertex of which represents black *A*. The intersection of these two conical surfaces is a circle *C–C–C*, representing the purest standards. Every cross section *F–F–F*, through the double cone, perpendicular to the axis, consists of concentric circles, each of which contains samples having equal contents of white, black, and color. These concentric circles are called isovalent circles. The standards in any one of these circles vary in all the standard colorimetric parameters, dominant wave-length, purity, and luminous reflectance. The quantities, in terms of which the various series of colors are presented in the Ostwald system, are in practice defined only by the actual samples and their arrangements in the published charts.

Ostwald arranged his full colors in a circle because hue is unquestionably a cyclical characteristic of color, having no extremes. The choice of a circle for the representation of the full colors of various hues was purely arbitrary, being the simplest manner of representing a cyclical variation in the absence of any other information which would make some other closed locus preferable. Ostwald, however, forgot the arbitrary and qualitative nature of the circular representation and deduced from it some rules of color mixture which might better have been investigated experimentally and used to provide the additional information necessary to discover the most appropriate shape of the locus representing the full colors of all hues. The results of color-mixture experiments indicate that the locus of Ostwald's full colors (semichromes) should have the shape indicated in Fig. 12, or some perspective view of that shape, from any point of view that may be chosen. Such a perspective view is called a projective transformation, and the mathematics of such transformations have been discussed in Chapter VII. The locus shown in Fig. 12, and appropriate for the representation of the color-mixture behavior of Ostwald's full colors, cannot be transformed into a circle by any projective transformation, or any perspective view.

Some of the rules of color mixture which Ostwald derived from his circle construction, and which he employed in the determination of intermediate hue steps and scales of color content are inconsistent with his fundamental definitions of those quantities. In particular, the presence of a sample labeled *pa* in every one of the hue charts is misleading and erroneous if judged on

[879] K. W. F. Kohlrausch, Physik. Zeits. **21**, 396, 423 (1920).
[880] A. König and C. Dieterici, Zeits. f. Psychol. u. Physiol. d. Sinnes. **4**, 241 (1892); also A. König, *Gesammelte Abhandlungen* (Barth, Leipzig, 1903), p. 214.
[881] S. Rösch, Forts. Miner. Krist. u. Petrogr. **13**, 73 (1928).
[882] H. E. Ives, J. Frank. Inst. **180**, 673 (1915); **195**, 23 (1923).
[883] O. Meissner, Zeits. f. Physik **21**, 68 (1924).
[884] E. G. Jacobson, More Business **2**, 2 (1937).
[885] F. Birren, *Color Dimensions* (Crimson Press, Chicago, 1934).

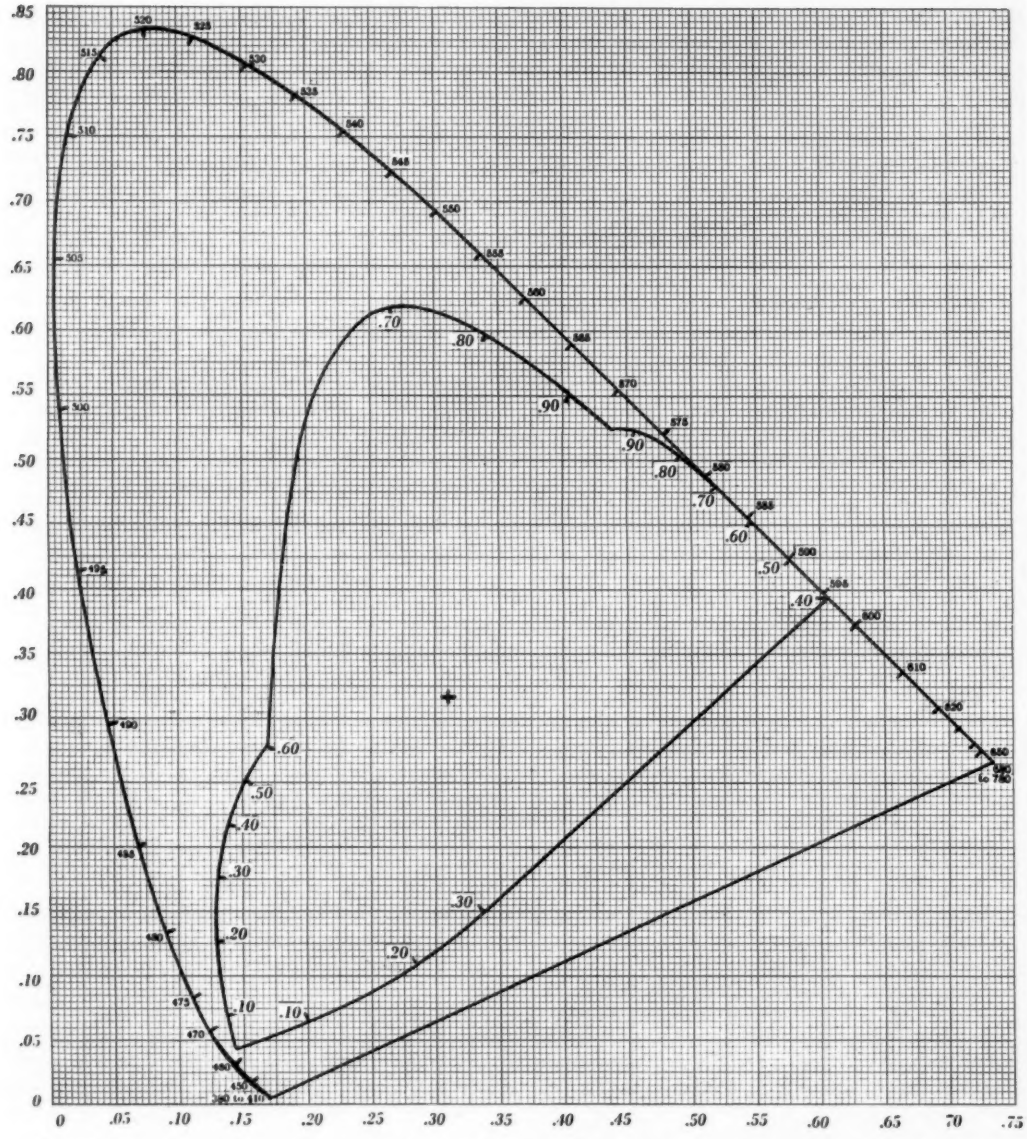


FIG. 12. Ostwald's semichromes represented in I.C.I. chromaticity diagram. Luminous reflectances are indicated.

the basis of Ostwald's definitions of full color and color content. Although surfaces can be secured which are properly labeled *pa* for the yellow, orange, red, and red-purple hues, such close approaches to the full colors cannot be attained with any available dyes or pigments for any of the blue-purples, blues, or greens. This fact, as well as the facts of color mixture referred to in the discussion of the appropriate locus for the full colors, indicates clearly that the purest standards contained in the Ostwald charts should not be arranged in a circle, and that the solid representing the Ostwald system of standards should not be a symmetrical conical figure. The

behavior of the purest Ostwald standards in disk (or other additive) mixtures can be deduced correctly from graphical constructions only if the figure representing the standards is some projective transformation of the locus of the corresponding points in the I.C.I. diagram. Ostwald rejected the work of Maxwell, Grassman, Helmholtz, König, and Ives, and thereby deprived himself of the quantitative data and methods necessary to perfect his system. Other investigators[873, 878] have endeavored to harmonize Ostwald's system with the quantitative results of color mixture experiments but the disciples of Ostwald have shown little interest in these analyses.

The Ostwald system, even in the idealized form derived by Richter,[878] does not include all possible colors of the light reflected by surfaces.[717, 740, 862, 881, 886, 887] The Ostwald system does not recognize and has no manner of specifying any colors which are represented on the I.C.I. diagram outside the locus of the full colors (Ostwald's semichromes) shown in Fig. 12. Surfaces having purities greater than those of Ostwald's full colors can be produced with available pigments and dyes. Figures 20 (for Illuminant *C*) and 21 (for Illuminant *A*) of Chapter VII show the possible purities of light reflected by surfaces having the luminous reflectances indicated,[740] and indicate the large gamut of possible colors outside the locus of the Ostwald full colors, and therefore not representable in the Ostwald system. They might be specified in terms of negative white contents, but negative values cannot be included in the geometric series, in terms of which isotint series of colors are defined. Therefore, although surfaces having purities greater than Ostwald's semichromes are obtainable with available dyes and pigments, they cannot be represented on the standard (logarithmic) Ostwald charts, or either inside or outside the Ostwald color solid. Similarly, reflecting surfaces having conceivable spectrophotometric characteristics[717, 862] can have greater purities for any specified luminous reflectance than can be matched with any conceivable mixture of a full color and a perfect white. These surfaces might be specified in terms of negative black contents, but they cannot be represented on any of the logarithmic Ostwald charts, nor in any conceivable spatial relationship to the color solid constructed from those charts. As a matter of fact, the only colored surfaces of maximum possible purities for given reflectances which can be represented in the Ostwald system are the full colors (Ostwald's semichromes), which have the greatest efficiencies (moments, or tinting strengths) in disk mixtures. Otherwise the ideal Ostwald color solid corresponds to only a portion of the interior of the solid representing conceivable colors reflected by surfaces. In this respect the Ostwald system is inherently more limited than the Munsell system, which neither postulates nor implies any limits to the attainable chromas, but can be extended in principle to include the colors of all surfaces up to and including the limits which are imposed by psychophysical facts.[834, 861]

Ostwald attributed psychological significance to the arrangements of colors in his system. His claim is based primarily on the use of the logarithmic rule for the subdivision of the scales, but the applications of this rule, especially in the mixtures of white with the full colors, have little or no psychological significance or justification. Noticeability of color differences in the various series was subordinated to the effort to fit the colors into a regular and symmetrical scheme of representation. The Ostwald system is essentially an empirical psychophysical system, specifying color in its own arbitrary scales, which were derived from additive disk mixtures by the use of inaccurate and in some cases erroneous rules of color mixture, which correlate suggestively but not accurately with the behaviors of pigment mixtures, and which are not related closely to any of the psychological attributes of color sensation except hue.

14. COLOR DICTIONARIES

The widespread use of color names in all lines of activity, commercial and industrial, as well as artistic and literary, has led to the compilation of color dictionaries. The purpose of these dictionaries has been to establish specific names for definite colors and to provide ready means of

[886] N. Nyberg, Zeits. f. Physik **52**, 406 (1928).
[887] N. T. Federov, Light Engineering U. S. S. R. No. 1, 11 (1937).

determining the color properly associated with any established name. The function of a color dictionary is accomplished by the provision of standards which, when observed under the proper conditions, yield the colors associated with various names. Usually the standards are arranged in some systematic manner which facilitates the search for the standard, and thus the name most nearly representative of any particular sample.

15. RIDGWAY COLOR DICTIONARY

Robert Ridgway, Curator of the Department of Birds of the United States Museum, published

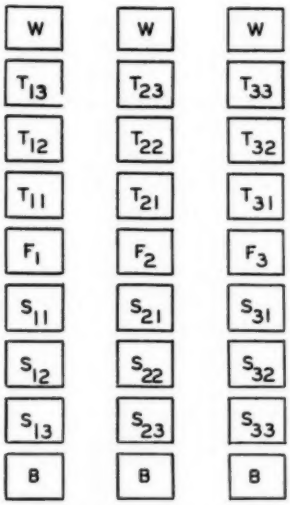


FIG. 13. Ridgway tints and shades of full colors.

a color dictionary in 1886, and a much enlarged edition in 1912.[888] The 1912 edition is still widely used for the description of the colors of birds, flowers, and insects. Fifty-three pages contain a total of 1113 different painted samples. Each of these samples is named and also identified according to a notation corresponding to the systematic groupings of the standards on the charts. The color names are also listed alphabetically in a table which gives the notation of each color so that it can be found easily in the charts containing the painted standards. The standards contained in the charts are painted reproductions of the colors obtained by a systematic series of additive mixtures obtained by the use of the Maxwell spinning disk. Sectors of white, black, a standard neutral gray, and chromatic papers of the highest obtainable purities for thirty-six approximately uniformly spaced hues were employed for the synthesis of the original colors, of which the standards in the Ridgway dictionary are painted reproductions. The thirty-six chromatic papers gave colors which Ridgway called *full colors*.

The first group of twelve charts, each representing three neighboring hues, contain the "tints" and "shades" of the *full colors*. Sectors of each of the full colors, occupying 90.5 percent, 77.5 percent, and 55 percent of the area of the spinning disk, were used with the remaining sectors of the disk covered with white to produce three tints of each of the full colors. Similarly, sectors of the full colors, occupying 55 percent, 29.5 percent, and 12.5 percent of the disk, were used with sectors of black paper over the remaining area of the spinning disk to produce three shades of each full color. These tints, T_{11}, T_{12}, $\cdots T_{33}$, and shades, S_{11}, S_{12}, $\cdots S_{33}$, are placed in order, Fig. 13, above and below the corresponding full color, F_1, F_2, F_3, in a column at the top of which was placed a white sample W and at the bottom of which was placed a black sample B. The thirty-six full colors are represented by the central sample F of thirty-six columns in the first twelve charts, with the tints T of each arranged from the full color F to white W at the top of the column, and the shades S arranged in order from the full color F to black B at the bottom.

For the determination of the colors represented in a second group of charts the sector formerly occupied by the full color was reduced to 68 percent of its area and the remaining 32 percent was occupied by a standard neutral gray material. This additive combination of a chromatic sample with a neutral gray was called a *broken color* and the tints and shades of this broken color were established by use of the same sectors of black and of white as were used to produce the tints and shades of the full colors. These broken colors and their tints and shades are arranged between white and black samples in

[888] R. Ridgway, *A Nomenclature of Colors for Naturalists* (Little, Brown and Company, Boston, 1886); *Color Standards and Color Nomenclature* (Washington, D. C., 1912).

thirty-six columns in a manner exactly analogous to those representing the tints and shades of the full colors. Each color represented on a third group of charts was established by reducing the sector occupied by the full color to 42 percent of its original area and filling the remaining portion of that area with the standard neutral gray. Similarly, each color represented on a fourth group of charts was obtained by reducing the sector occupied by the full color to 23 percent of its original area, and substituting neutral gray for the remainder of that sector. In a fifth and a sixth group of charts the sector occupied by the full color was reduced to 10 percent and 4.5 percent of the area occupied by the full color in the synthesis of the corresponding tint or shade of the full color.

Ridgway's nomenclature, e.g., "full color," "tint," and "shade," is strikingly similar to that of Ostwald, and the scales of tints and shades of the full colors have almost identical structures in the Ridgway and the Ostwald systems. The scales of tints and shades of the broken colors are not comparable with any of the scales of the Ostwald system.

No spectrophotometric analysis of the Ridgway system of colors has been published and consequently conversion from I.C.I. to Ridgway specifications, or the reverse, cannot be accomplished at present. However, the Ridgway system was not primarily intended to provide a standard system of color measurement. Its primary purpose, to establish definite meanings for color names for use in the natural sciences, was well accomplished. The arrangements and systematic notations adopted by Ridgway serve the purpose of ready reference, which was the original intent, and in this respect their admitted failure to conform with the attributes of color sensation is unimportant.

16. MAERZ AND PAUL DICTIONARY OF COLOR

The foremost authority for color names in general use is the Maerz and Paul *Dictionary of Color*.[889] The usage indicated in this dictionary is based on a thorough survey of all the available information concerning color names. Over seven thousand different color standards were printed under carefully controlled conditions, and after the printed pages were complete about four thousand color names were keyed to the appropriate standards according to the concensus of prevalent usage. The arrangement of the standards on the pages was designed for convenience and accuracy in printing, and no attempt was made to follow any color system or to propose a new system based on the arrangement of the standards in the dictionary. The arrangement is convenient for the purposes of the dictionary, all related colors being grouped together systematically. Eight chromatic pigments were used for the production of the standards. The hues of these pigments were selected so that all hues could be produced with high purity by superimposed impressions of pairs of pigments with neighboring hues. On most of the pages, such as that illustrated in Fig. 14, eleven small rectangles are printed with one chromatic pigment in a series of increasing strength across the top of the page, beginning with an unprinted rectangle *A* at the upper left corner. A similar series of eleven rectangles is printed with a pigment of neighboring hue in a column at the left side of the page increasing in strength toward the bottom of the page, beginning with the unprinted rectangle *A*. The same strength of the first pigment is printed in eleven rectangles in the column below each of those in the top row, and the same strength of the second pigment is printed in the eleven rectangles to the right of each of those in the left-hand column. The resultant two-pigment combinations exhibit a representative selection of hues intermediate between those of the two pigments used, with purities varying from zero to the maximum obtainable with well-selected pigments. Similar combinations of these same pigments are printed over gray impressions having several different reflectances, to give the desired variation of luminous reflectance.

17. SPECIAL COLOR STANDARDS

The Textile Color Card Association of the United States issues seasonal and standard color cards[890] showing the seasonal and standard colors

[889] A. Maerz and M. R. Paul, *A Dictionary of Color* (McGraw-Hill Book Company, Inc., New York, 1930).

[890] Textile Color Card Association of the United States, Inc., *Standard Color Card, Spring Color Card, Fall Color Card, U. S. Army Color Card* (New York).

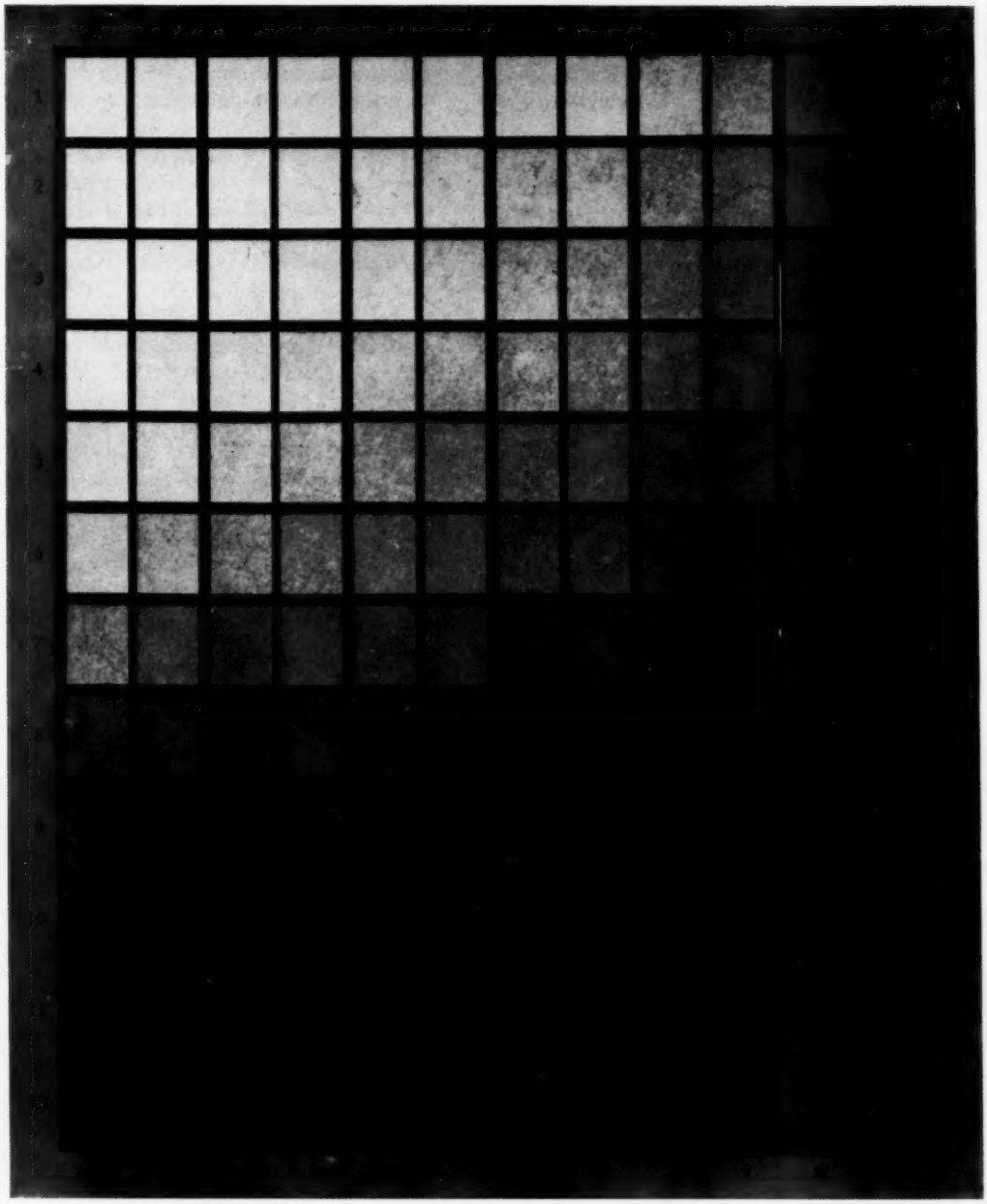


FIG. 14. Photograph of typical page of Maerz and Paul, *Dictionary of Color*.

associated with specific names which are generally accepted in the American textile industry. There are 216 samples in the 1941 edition of the Standard Color Card. These are of satin-backed crepe about $2\frac{1}{2}$ inches square, folded to show both satin and crepe surfaces. The Association

expects to secure spectrophotometric data for these samples and to publish I.C.I. colorimetric specifications for them, but this program has been delayed by the war. The British Colour Council *Dictionary of Colour Standards*,[891] which serves a similar purpose for the textile and related industries of the British Empire, consists of two volumes. One volume contains a series of ribbon samples, satin and grosgrain in each, of 220 named and coded standards. The second volume is a dictionary containing the names, the sources of the colorants and names, readings made on the samples with the Lovibond tintometer, and I.C.I. specifications based on measurements made with the Guild colorimeter, employing Illuminant *B*.

The *Repertoire de Couleurs*[892] of the Société Française des Chrysanthémistes, and the *Horticultural Colour Chart*[893] of the Royal Horticultural Society are in current use in horticulture and related fields.

18. SYSTEMATIC COLOR NAMES

A systematic method of naming colors has been devised and recommended by the Inter-Society Color Council. These names have been adopted by the American Pharmaceutical Association and the U. S. Pharmacopoeial Convention for use in the National Formulary and the U. S. Pharmacopoeia. The details of this method were developed at the National Bureau of Standards,[894] and the method is generally referred to as the I.S.C.C.–N.B.S. method of designating colors. This plan provides a systematic name for every color perceived as belonging to a surface, but the number of designations is small (slightly over three hundred) in the interests of simplicity. Consequently, many slightly different colors are assigned the same name by this method. Commonly understood names are employed, with modifiers which are familiar and clearly defined in the system. The boundaries between groups of colors designated by each name correspond as closely as practical to common usage, and have been expressed in terms of the Munsell system.[895] Central notations for these names have also been published,[896] and samples representing the central colors for 56 of the names have been prepared for use as standards for the naming of the colors of soil.[897] A chromaticity diagram has been published showing a related system for naming colored lights.[898]

19. AMERICAN EMERGENCY STANDARD

An emergency standard for the specification and description of color has been adopted by the American Standards Association. The provisions of this standard can be summarized as follows:[899]

1. The spectrophotometer is the basic instrument in the standardization of color.

2. The basic specifications of color are computed from spectrophotometric data by the use of I.C.I. (1931) standard observer and (preferably) Illuminant *C*, and expressed in terms of *Y*, *x*, and *y*, or dominant wave-length Λ, and excitation purity.

3. The *Munsell Book of Color*[852] is recommended for the popular identification of color and the idealized Munsell system[834] for the precise interpretation of basic (I.C.I.) specifications.

4. Descriptive color names according to the I.S.C.C.–N.B.S. system are recommended for use whenever general comprehensibility is desired and precision is not important. The use of color names is not recommended for color specifications.

---

[891] British Colour Council, *Dictionary of Colour Standards* (London, 1934), two volumes.

[892] Société Française des Chrysanthémistes, and R. Oberthur, *Repertoire de Couleurs pour Aider à la Determination des Couleurs des Fleurs, des Feuillages et des Fruits*, (82 pp., 365 color plates) Rennes, Impr. Oberthur (Paris, Libraire horticole, 1905).

[893] British Colour Council, *Horticultural Colour Chart* (London, 1938).

[894] D. B. Judd and K. L. Kelly, J. Research Nat. Bur. Stand. **23**, 355 (1939).

[895] D. Nickerson, 1940 A. S. T. M. Standards on Textile Materials, D-13 Manual; also Am. Dyestuff Reporter **29**, 392 (1940).

[896] D. Nickerson and S. M. Newhall, J. Opt. Soc. Am. **31**, 587 (1941).

[897] T. D. Rice, D. Nickerson, A. M. O'Neal, and J. Thorp, U. S. Dept. Agr. Misc. Publ. 425 (1941).

[898] K. L. Kelly, J. Opt. Soc. Am. **33**, 627 (1943).

[899] American Standards Association, Z-44, J. Opt. Soc. Am. **32**, 694 (1942).

A Method of Growing Oriented Sections of Certain Optical Crystals

C. D. WEST
Polaroid Corporation, Cambridge, Massachusetts
(Received September 25, 1944)

A new method is described for growing extended crystal sections normal to a threefold axis from melts of sodium nitrate, and of certain alkali halides such as potassium bromide and potassium iodide. Such sections result when crystallization in the melt is started at a mica cleavage surface and is allowed to proceed in a direction normal thereto, the mica acting as an insoluble, infusible seed for the new crystal growth. The mica cleavage surface adsorbs from the melt a hexagonal net of alkali ions whose side is 5.17A, and this plane net constitutes the first layer of the new crystal growth.

I. SODIUM NITRATE

FEUSSNER[1] first showed how polarizing prisms of desirable properties could be prepared from thin sections of crystals having high birefringence; in particular, sodium nitrate, whose refractive indices for sodium light are $\omega=1.587$, $\epsilon=1.336$, $\omega-\epsilon=0.251$. Feussner's prisms give normal polarized fields, and they are, for uniaxial negative crystals, ω-ray prisms, as contrasted to Nicol and other calcite prisms as ordinarily constructed, which are ϵ-ray prisms. The polarized images transmitted by the two types are, respectively, anastigmatic and astigmatic; and this property of ϵ-ray prisms is troublesome in image-forming systems such as polarizing microscopes. On the other hand, the ϵ ray is transmitted over a wider spectral range than the ω ray for both calcite and sodium nitrate.

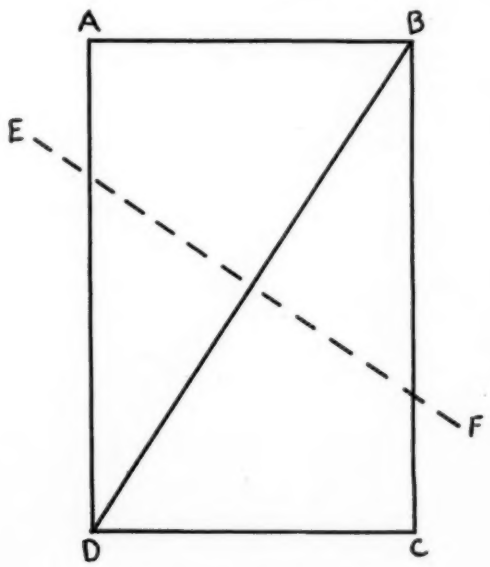


FIG. 1. Orientation of the crystal wafer.

As for the orientation of the crystal wafer (Fig. 1), an adequately explicit statement is found in Feussner's original paper; a random section of the crystal (trace *B*, *D*) is oriented so that its optic axis falls in the plane *E*, *F*, which is defined as perpendicular both to the plane of the section *B*, *D* and to the plane of incidence *A*, *B*, *C*, *D*.

Optically, one section is as good as another. Mechanically, a section perpendicular to the axis has equal thermal expansion in all directions in its plane, whereas a section of calcite or sodium nitrate parallel to the axis shows a marked anisotropy of thermal expansion.

Since Feussner's time there have been repeated attempts to grow single crystals of sodium nitrate from which polarizing prisms could be prepared. In Germany, L. Wulff[2] described his work on growing such crystals from water solution; Stöber[3] wrote several papers on growing sodium nitrate from the melt. In Russia, Tzekhnovitzer[4] continued with Stöber's methods. This crystal is also mentioned by Kremers.[5]

[1] K. Feussner, Zeits. f. Instrumentenk. **4**, 41–50 (1884).

[2] L. Wulff, Sitz. Preuss. Akad. Wiss. **135**, 879 (1896).
[3] P. Stöber, Zeits. f. Krist. **61**, 299–317 (1924); Neues Jahrb. Mineral. **A57**, 139–62 (1928); Chem. Erde **6**, 357–67; 453–67 (1930).
[4] E. Tzekhnovitzer, J. Phys. Chem. U.S.S.R. **5**, 1452–8 (1934).
[5] H. Kremers, Ind. Eng. Chem. **32**, 1478–83 (1940).

We have found, what was not previously reported in the literature, that if crystallization in a sodium nitrate melt is started on a mica cleavage surface, the sodium nitrate crystal is perfectly oriented by the mica with a basal plane parallel to the mica cleavage. There then remains only to control the flow of heat so that the entire melt is converted into a single crystal; and finally to cool the crystal so formed to room temperature. At the end of the process, on undertaking to strip the mica from the surface of the crystal, it is found to adhere quite strongly, enough so to cause frequent glide twinning lamellae. The stripped sodium nitrate surface reproduces exactly the mirror qualities of the mica surface from which it grew. This process succeeds with the several different varieties of mica we have tried so far, notably commercial grades of muscovite and phlogopite. With careful handling, the same piece of mica may be used repeatedly in this process. We have grown areas of sodium nitrate $7\frac{1}{2}$ in.$\times$15 in.$\times\frac{3}{4}$ in. thick without difficulty (Fig. 2). It is of particular interest that sodium nitrate can be grown from a floating piece of mica downwards, or in the opposite direction from that which Stöber and later workers considered essential. Even more surprising is the observation that mica sheets of density 2.75 float readily on molten sodium nitrate of density 1.89; the explanation is that the mica is not readily wetted by the melt.

To judge from the literature, the present method is the simplest one for growing single crystals of considerable size so far described.

We have here an example of what is called the mutual orientation of crystal species. In effect the mica acts as an insoluble infusible seed for the growth of an alien crystal. Although this phenomenon has been described extensively in minerals, and in the deposition of crystals of one kind from solution on the surface of another species in the laboratory,[6] the corresponding processes of growth from the melt have apparently been very little studied. Following is a review of present knowledge of the structures in question, and their thermal properties.

[6] L. Royer, Bull. soc. franç. min. **51**, 7–159 (1928); J. Willems, Zeits. f. Krist. **105**, 53–68 (1943); A. Neuhaus, Zeits. f. physik. Chemie **192**, 309–31 (1943); C. Sloat and A. Menzies, J. Phys. Chem. **35**, 2005–21 (1931).

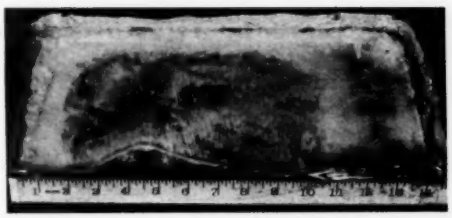


FIG. 2. Slab of $NaNO_3$ grown under mica from a five-pound melt (inch scale for comparison).

Structure of the Cleavage Surface of Mica

The cleavage surface of mica is well known to be a plane net of oxygen atoms, which are the oxygen bases of SiO_4 tetrahedrons linked through their corners. There are in the net only $\frac{3}{4}$ the number of oxygen atoms that would be required to give a simple triangular net; the hole left by the missing fourth oxygen atom is occupied, prior to cleavage, by potassium atoms which bind the two identical oxygen nets together prior to cleavage. After cleavage, one half of the holes on each surface is occupied in a statistical manner by the potassium atoms.[7]

It is also recognized that these surface potassium atoms are readily exchanged for other cations, such as sodium, in water solution.[8] There is little doubt that the same exchange takes place between the Na of molten $NaNO_3$ and the K of cleaved mica.

Prior to cleavage, and after cleavage so far as is known, the side of the hexagonal net in muscovite has a length of 5.17A.[9] The thermal expansion of two micas in the cleavage plane was measured by H. Ebert[10] who gives $\Delta l/l_0$ in mm/m starting at 0° as 2.65 and 4.15 at 300°, 5.05 and 6.90 at 500°, respectively, for muscovite and phlogopite.

Structure of $NaNO_3$

Kracek[11] found that the $NaNO_3$ crystal on heating undergoes a very interesting gradual transition ending at 275°, or 35° below the

[7] G. Scheibe, Zeits. f. angew. Chemie **52**, 635 (1939).
[8] U. Hofman and W. Bilke, Kolloid Zeits. **77**, 249–50 (1936).
[9] W. Jackson and J. West, Zeits. f. Krist. **76**, 211–27 (1931); **85**, 160–4 (1933).
[10] H. Ebert, Physik. Zeits. **36**, 258 (1935).
[11] F. Kracek, J. Am. Chem. Soc. **53**, 2609–24 (1931).

melting point. Kracek, Posnjak, and Hendricks[12] in an x-ray study found that the oxygen lattice by itself no longer gives crystal scattering at this temperature, and proposed in explanation for this a rotation of the NO_3 ions in their planes about the threefold axes on which they lie. Such a structure would, however, require smaller Na−O and especially O−O distances than are found in crystals of this type, namely, Na−O = 2.22A, O−O = 2.53A, as against expected distances of 2.44 and 3.1A, respectively.[13] It appears that before a final exact picture can be given of the $NaNO_3$ structure above 275°, which is apparently the same as that of KNO_3 above 128°,[14] the available diffraction intensity data will require a more elaborate analysis than has so far been given.

There are three measurements of the thermal expansion of $NaNO_3$, two by the x-ray powder method,[12,15] and one by interferometer.[16] These data are summarized in Table I. These foregoing figures show that the total contractions of both the mica in its cleavage plane and the $NaNO_3$ crystal perpendicular to its axis are small and at least of the same order.

Mechanism of Mutual Orientation

On placing the mica sheet in contact with the $NaNO_3$ melt, it is reasonable to suppose that the Na ions exchange with the K ions occupying half of the holes, and themselves occupy the remaining vacant holes. We then have a plane triangular net of Na ions, of side $5.17(1+0.003) = 5.19$A. Now this differs only by $\frac{5.19-5.09}{5.09} = 2$ percent from the plane net of Na ions in the $NaNO_3$ crystal at its melting point. It is therefore only natural, on cooling the melt below the melting point, for the second layer, namely, rotating NO_3 ions, to take up their prescribed positions on top of the layer of Na ions. This they can do in one of two ways, differing by a rotation of the $NaNO_3$ crystal through 60° or 180°. The third layer, of Na atoms, can go on in only one way, and the same is true of succeeding alternate layers of NO_3 and Na ions.

If crystallization of $NaNO_3$ starts at only one point on the mica, one would expect to obtain a corresponding single crystal of $NaNO_3$; but if crystallization starts simultaneously at several points on the mica, then the probability would be quite large of having trigonal axis twinning, as a result of the optional orientation of the second atomic layer of the crystal consisting of NO_3 ions. In practice, trigonal axis twinning is observed fairly commonly, recognized either by interruption of a cleavage, or as an irregular boundary surface, perpendicular to the plane of the crystal disk, between different areas on the disk; parting along this boundary is quite easy.

Quite often, too, small sharply limited areas in the crystal disk are found to be inclined at an appreciable angle to the optic axis; however, the crystallographic orientation of these areas has not been exactly determined as yet. Only rarely have we obtained a whole single crystal disk inclined to the optic axis; when such a disk cools it has a markedly cylindrical surface, as might be expected.

TABLE I.

	A+P[a]	KPH[b]	S+M[c]
a_{25}	—	5.064	5.0600
a_{280}	—	5.082	5.0823*
$\Delta a/a_{25}$ mm/m	2.77	3.670	4.4000
c_{25}	—	16.82	16.784
c_{280}	—	17.61	17.448*
$\Delta c/c_{25}$ mm/m	45.5	47.00	39.500

[a] See reference 16.
[b] See reference 12.
[c] See reference 15.
* Extrapolation from 220°.

Crystallographically the orientation is: (111) of $NaNO_3$ parallel to (001) of mica; $[2\bar{1}\bar{1}]$ direction of $NaNO_3$ parallel to b axis or twofold axis [010] of mica, which in biaxial muscovites is the same as the γ optical direction or trace of plane of optic axes. The twofold axis $[1\bar{1}0]$ of $NaNO_3$, whose direction is shown by the trace of a cleavage plane on (111), is thus always parallel to the β optical direction which is the a axis [100] of the mica. The alternative orientations are $[1\bar{1}0]$ and $[\bar{1}10]$ of $NaNO_3$ coincident with [100] of mica. The structures are represented in Fig. 3.

[12] Kracek, Posnjak, and Hendricks, J. Am. Chem. Soc. **53**, 3339–48 (1931).
[13] W. H. Zachariasen, Zeits. f. Krist. **80**, 137–53 (1931).
[14] Kracek, Barth, and Ksanda, Phys. Rev. **40**, 1034A (1932.)
[15] H. Saini and A. Mercier, Helv. Phys. Acta **7**, 267–72 (1934).
[16] J. Austin and R. Pierce, J. Am. Chem. Soc. **55**, 661–68 (1933).

Variation of Composition of the Melt

The effect of impurities in the melt is readily studied by this method. Sodium nitrate will tolerate quite large percentages of $AgNO_3$, for example, over $8\frac{1}{2}$ percent by weight, and smaller amounts of NaI in the melt, and still give single mixed crystals. The bottom layers of such crystal disks are decidedly granular in appearance, which has the effect of scattering the transmitted light. In the present process we have used Baker's analyzed sodium nitrate and Mallinckrodt sodium nitrate analytical reagent without further purification, and are inclined to doubt that such purification would have any perceptible effect.

Properties of Sodium Nitrate Monocrystals

We have observed that these crystals have a pronounced thermoplasticity; when immersed in a bath of Wood's metal, they are readily bent to a small radius of curvature at a temperature as low as 170°. Even slight plastic distortions are easy to observe by optical methods using polarized light, especially when the sections perpendicular to the axis are examined in a conoscope. The common distortion is a wandering of the optic axis with reference to a fixed plane, which is accounted for in terms of simple bending. We have not observed optical biaxiality in badly deformed crystals at all. Closer examination reveals a slight convexity (spherical, cylindrical, or sphero-cylindrical) of the surface of the crystal disk which grew from the mica, evidently owing to the method of cooling employed. Under favorable conditions the radius of curvature is of the order of four to eight meters. It will be recalled that when an ordinary melt solidifies under similar conditions, the surface always assumes a concave curvature as a result of thermal contraction.

Another property of these crystal sections is their very notable resistance to thermal shock. This we estimate to be of the same order or better than that of ordinary glass under identical conditions. The crystals can be quenched in oil from just below the melting point to room temperature or lower without cracking.

It is possible to work these crystals on the lathe without difficulty. When the basal plane is faced off with a tool point, a pattern of threefold symmetry results, which shows how the hardness varies with direction on this face. A similar pattern is readily developed on the basal plane of calcite, while a fourfold pattern can be turned on a rocksalt cleavage face. It is very easy to introduce twinning lamellae in the mechanical operations of shaping sodium nitrate crystals.

The refractive indices of a polished basal section as measured on an Abbe refractometer are $\omega_D = 1.5852$, $\epsilon_D = 1.3350$, $(\omega - \epsilon)_D = 0.2502$ in agreement with previous values. In the mixed crystal with $8\frac{1}{2}$ percent $AgNO_3$ already referred to, these are increased by 0.01 to $\omega_D = 1.5955$, $\epsilon_D = 1.3431$, $(\omega - \epsilon)_D = 0.2524$.

II. ALKALI HALIDES

In a preliminary survey it was found that KI, RbI, NaI, and KBr are readily oriented by crystallizing from the melt on a mica surface, with (111) parallel to (001) of mica, and with a twofold axis $[1\bar{1}0]$ or $[\bar{1}10]$ coincident with [100] of mica as before. These all crystallize from the melt with the NaCl structure; the translations $[1\bar{1}0]$ are, respectively, at room temperature $a/\sqrt{2} = 4.99$, 5.18, 4.57, 4.65A.

Eucken and Dannöhl[17] measured the linear expansion coefficient $\alpha(t)$ for KBr 45.0 to 680.9°, and for KI 43.7 to 607.8°. By integrating their equations we calculate the total contractions from 600 to 0° in mm/m as 28.8 and 29.8, respectively; the corresponding values for the range 700° to 0° are 35.7 and 37.5, which must be several times the contraction of mica in its cleavage plane over the same intervals. By increasing $a/\sqrt{2}$ by 3.6 percent we estimate the hexagonal nets of KBr and KI to have lengths of 4.81 and 5.16A at the respective melting points.

We have grown crystals, in the form of disks parallel to (111) of dimensions 40 mm × 4 mm thick, of KI and KBr. The whole discussion under $NaNO_3$ applies to these save that owing to the optical isotropy of the alkali halides the polariscopic examination shows only the condition of strain, not of crystal orientation, in the monocrystal disk. Only the cleavage fissures on the cube faces show the orientation of these disks.

The alkali halide crystals also withstand quite drastic quenching. A cleavage block of NaCl, 28 × 16 × 6 mm, quenched from about 700° in oil

[17] A. Eucken and W. Dannöhl, Zeits. f. Elektrochemie **40**, 819 (1934).

at 60°, exhibits a symmetrical pattern between crossed polarizers similar to that which is observed in a glass block treated in the same manner.

Among a number of other compounds crystallized from the melt on mica cleavage surfaces, we find that $CsNO_3$ and $CdCl_2$, both crystals with trigonal symmetry, give well-oriented growths, whose laws have not yet been established.

Experimental

The sodium nitrate crystal basal disks have successfully been grown and cooled without cracking in shallow flat-bottomed cups of aluminum of various thickness from 0.002″ to 0.012″ (0.05 to 0.3 mm). Mostly the melts have been made in an electric muffle oven, in which an upward flow of heat can be provided by leaving the door slightly ajar. A block of insulating material is shaped with a well or depression closely fitting the diameter and depth of the aluminum cup, and providing about one-inch wall thickness around the cup. Two procedures are employed: growth from the top down from a floating disk of mica; and growth from the bottom up, from a disk of mica sunk in the melt. The second procedure is generally preferred as it allows for escape of gas bubbles, and permits the large c-axis contraction in the simplest manner and thus minimizes thermoplastic distortion; however, there are no gross differences in the products prepared by these two methods.

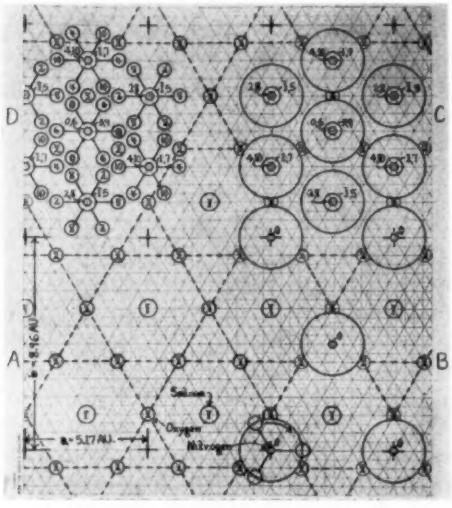


FIG. 3. Steps in the growth of $NaNO_3$ crystal from the melt on a mica cleavage surface. *A*. Mica cleavage surface occupied by Na at height $\bar{1}$. *B*. First layer of NO_3 added at height 0. *C*. $NaNO_3$ on mica, NO_3 rotating. *D*. The same, NO_3 fixed. Heights are indicated as twelfths of the c axis (see Table I for this axis).

In the first method the cup in the well is charged and melted in the oven, the mica disk is floated on without entrapping air bubbles, the door is closed and after an interval the electric current is turned off. When cool, the aluminum and mica are stripped from the single crystal disk. In some large scale experiments on the growth of crystal disks 3″ in diameter by this method, it was found that the undesirable warping of the disks which has already been referred to could be largely eliminated by the presence of a glass disk about $\frac{1}{16}$″ thick at the bottom of the cup.

In the second method the cup is placed on a block of heat-conducting material such as aluminum, the charge is melted and the mica disk is sunk therein, the preheated insulating well is inverted over the cup, the oven door is closed, and the electric current is turned off. The conducting base here provides for a flow of heat from the bottom of the cup as desired.

Single crystals have been grown in very short intervals by these two methods. Too rapid cooling is observed to result in a turbid crystal, a condition which is also encountered in the manufacture of artificial ice. When the bottom of an aluminum cup charged with sodium nitrate melt containing a sunken disk of mica is drastically cooled by contact with mercury or ice, oriented single crystals of appreciable area result. In this connection we refer to the interesting experiments of Stöber[3] from which he concluded that a prismatic crystal grows the faster, under a given set of conditions, the smaller its cross section; in particular for the growth of sodium nitrate parallel to the c axis, a section one sq. cm in area requires 0.77 minute to grow 1 mm, while a section of 95 sq. cm requires 60 minutes to grow the same distance.

We have also observed oriented growths of sodium nitrate from the melt on polished basal surfaces of calcite; but since it does not seem to be possible either to cool the sodium nitrate or

detach it from the calcite without cracking, this is of little practical interest. Johnsen[18] previously described the oriented growth of sodium nitrate from the melt on a calcite cleavage surface.

Since the alkali halide melts wet mica readily, it is found easiest to use a Hotcone heater, and to support the mica disk, about 0.001″ thick, by three fine copper wires from a glass rod, whose height can be readily adjusted by hand. The mica disk is lowered until it just touches the surface of the melt, which is in a porcelain dish of somewhat greater radius than the mica disk. The temperature is lowered so that a single crystal disk grows downward from the mica to a suitable thickness, this is then transferred to an oven in which it can be cooled slowly. The mica may be stripped from the crystal while it is still hot; if it is left adhering to the crystal there is more danger of cracking. The disk acquires a pronounced convexity on the mica side while cooling, ending up in the form of a meniscus lens, and the mica if left adhering often breaks up under the large stresses to which it is subjected.

In conclusion I wish to thank my associate, Mr. Frederick J. Binda, and others who have generously cooperated in the prosecution of this research, and Mr. E. H. Land for his continued interest and support.

[18] A. Johnsen, Min. Petr. Mitt. **37**, 81 (1927).

Inter-Society Color Council Meeting

THE thirteenth annual meeting of the Inter-Society Color Council will be held at the Pennsylvania Hotel, New York City, on February 23 and 24, 1945.

The Distribution of Color Matchings Around a Color Center*

LUDWIK SILBERSTEIN AND DAVID L. MACADAM
Research Laboratories, Eastman Kodak Company, Rochester, New York

(Received October 2, 1944)

Experimental data concerning the accuracy of visual chromaticity matching (automatically constant luminance) which were previously summarized graphically, are compared with predictions based on the assumption of a normal frequency distribution in two dimensions. The standard deviations and correlation coefficient of the two conventional coordinates vary from point to point in the chromaticity diagram and values computed from the experimental data are presented. The major and minor axes, and the angle of inclination of the major axis of the ellipse representing the standard deviation of matchings in any specified direction from any central color, are computed from the parameters of the distribution function, as are the coefficients of the quadratic differential form representing noticeability of chromaticity difference. The conclusions previously published, based on graphical analysis of the data, are confirmed and strengthened by these results, eliminating the former dependence upon personal judgment in construction of ellipses through scattered points.

IN earlier papers[1,2] on the metrical properties of the two-dimensional manifold of equiluminous colors, the differential geometry of this manifold or of the "true color surface" representing it was derived from experimental results[3] concerning the standard deviation of color matchings in various directions around the central color. These results of numerous matchings were embodied in twenty-five oval curves surrounding as many centers spread over the color triangle, all of which turned out to be, nearly enough, ellipses, with axes varying in size and orientation from point (color center) to point. Thus, the element of distance or the line element of the manifold was expressed by a definite quadratic differential form of the stimuli-coordinates, $ds^2=g_{11}dx^2+2g_{12}dxdy+g_{22}dy^2$, with coefficients g_{ik} evaluated numerically from the size and orientation of those axes at the explored centers and, by interpolation, at intermediate points. The knowledge of these coefficients enabled us to compute the curvature of the manifold or of the representative true color surface at several points, and thus to show, at the conclusion of reference 1, that this surface differs emphatically from the plane or any developable surface, which, of course, is an intrinsic property of the twofold, wholly independent of the choice of coordinates.

The purpose of the present paper is to return to the experimental result, the elliptic distribution of the standard deviation around any chosen color center, and to show that it can be derived from general probability considerations with the aid of a comparatively simple and plausible assumption.

Let x, y be the stimuli-coordinates of a color match, and let $f(x, y)dxdy$ be the probability that a color match (for a given observer) shall fall within x to $x+dx$ and y to $y+dy$. In other words, let $f(x, y)$ be the frequency of points or color matchings around x_0, y_0 as the color center or point aimed at. *A priori*, we can say only that $f(x, y)$ will decrease with increasing $x-x_0$, $y-y_0$, and that its integral extended over the whole color triangle must be unity, $\iint f(x, y)dxdy=1$. This leaves the form of the function free within very wide limits. Let us now make the following assumption, first proposed by Cotes (1772) for the distribution of rifle shots on a target:

If any number n of color matchings, $x_1, y_1; \cdots x_n, y_n$, have been made, then *the most probable* point or color aimed at is *the center of gravity* of these colors, that is, the product of n factors $f(x_i-x_0, y_i-y_0)$ or the sum of n terms $\log f(x_i-x_0, y_i-y_0)$ attains a maximum when x_0, y_0 considered as variables are attributed the values

$$x_0=-\frac{1}{n}\sum x_i=\bar{x}, \quad y_0=-\frac{1}{n}\sum y_i=\bar{y}.$$

* Communication No. 995 from the Kodak Research Laboratories.

[1] L. Silberstein, J. Opt. Soc. Am. **33**, 1–10 (1943).
[2] D. L. MacAdam, J. Opt. Soc. Am. **33**, 18–26 (1943).
[3] D. L. MacAdam, J. Opt. Soc. Am. **32**, 247–274 (1942).

It can readily be shown* that $\log f$ must then be a quadratic function of x, y. Hence, with $u=x-\bar{x}$, $v=y-\bar{y}$ and a convenient notation for the coefficients,

$$f(x, y)=C\exp\left[-\left\{\frac{u^2}{\sigma_1{}^2}-\frac{2\rho}{\sigma_1\sigma_2}uv+\frac{v^2}{\sigma_2{}^2}\right\}\Big/2(1-\rho^2)\right], \quad (1)$$

where σ_1, σ_2, ρ, and C are independent of u, v, but may vary from center to center. This is the so-called "normal law" of frequency distribution for two variable attributes (u, v). If, in particular, $\rho=0$, then $f(u, v)$ is the product of $\exp(-u^2/2\sigma_1{}^2)$ and $\exp(-v^2/2\sigma_2{}^2)$, the familiar Gaussian or normal frequencies with standard deviations σ_1 and σ_2 of the two attributes taken separately. These attributes are then, and only then, not correlated with each other, whence the name *correlation coefficient* given to ρ by Bravais and others. (It is usually denoted by r.) The constant factor C is always to be evaluated in terms of σ_1, σ_2, ρ by means of the integral $\iint f(u, v)dudv=1$ extended over all possible values of u, v. If the twofold is unlimited and the integration of (1) is extended over u and v from $-\infty$ to ∞, the result is

$$C=\frac{1}{2\pi\sigma_1\sigma_2(1-\rho^2)^{\frac{1}{2}}}. \quad (2)$$

Since our twofold, the color triangle, is limited and finite, no such relation of course can hold rigorously. If the color center, however, is well inside the triangle, so that the values of u/σ_1 and v/σ_2 occurring in the integral exceed only a few units, positive and negative, then the expression (2) will give a very close approximation to the true value of C. If, on the other hand, the center chosen is in the spectrum or on the line of purples, C will differ from (2) substantially.

The three coefficients σ_1, σ_2, ρ themselves have a simple property, regardless of the extent of the whole twofold and of the position of the chosen center, namely, if any number n of points u, v, with $\bar{u}=\bar{v}=0$, have been observed, then *the most probable values* of these coefficients, for which, that is, the product of the n factors $f(u, v)$ attains a maximum, are given by

$$\sigma_1{}^2=\langle u^2\rangle_{Av}, \quad \sigma_2{}^2=\langle v^2\rangle_{Av}, \quad \rho=\frac{\langle uv\rangle_{Av}}{\sigma_1\sigma_2}, \quad (3)$$

where $\langle u^2\rangle_{Av}=(1/n)\sum u^2$, and so on. Since $\langle uv\rangle_{Av}{}^2 \leqq \langle u^2\rangle_{Av}\cdot\langle v^2\rangle_{Av}$, $\rho^2\leqq 1$, so that the quadratic form $u^2/\sigma_1{}^2+v^2/\sigma_2{}^2-2\rho uv/\sigma_1\sigma_2$ is always positive.

Let us now turn to the actual conditions of MacAdam's experimental color matchings.

In order to establish the match, the observer had to make only one adjustment of the instrument, which was deliberately set so that the locus of possible matches was a straight line (on the x, y chart), inclined at a fixed angle $\tilde{\omega}$ to the x axis. For every chosen angle $\tilde{\omega}$, at least fifty such adjustments were made and their average was taken as the correct reading for the color match. In other words, the observer was not free to choose the two attributes u and v, independently of each other, but only their resultant r, say, so that $u=r\cos\tilde{\omega}$, $v=r\sin\tilde{\omega}$ in a fixed direction and in the opposite direction, $\tilde{\omega}$ and $\tilde{\omega}+\pi$. This may be called a *guided* color matching, in distinction from a free one.

The element of area in polar coordinates is $rd\tilde{\omega}dr$. If we denote by $f(r, \tilde{\omega})d\tilde{\omega}dr$ the probability of a point falling within an angle $d\tilde{\omega}$ at a distance r to $r+dr$ from the center, then, by (1), for free color matches,

$$f(r, \tilde{\omega})=Cr\exp\left[-\varphi^2(\tilde{\omega})r^2\right], \quad (1')$$

where

$$\varphi^2(\tilde{\omega})=\left\{\frac{\cos^2\tilde{\omega}}{\sigma_1{}^2}-\frac{\rho\sin 2\tilde{\omega}}{\sigma_1\sigma_2}+\frac{\sin^2\tilde{\omega}}{\sigma_2{}^2}\right\}\Big/2(1-\rho^2). \quad (4)$$

Now, the probability that a free match shall fall exactly on a fixed line is manifestly nil, while a guided match is certain to fall somewhere on that line.

The actual experimental conditions, however, are well approximated by supposing that the guidance consists in constraining the matches to remain between two straight lines parallel to

* See, for instance, Bertrand's proof in Whittaker and Robinson's *Calculus of Observations* (Blackie and Son, London, 1924), p. 321.

TABLE I. Distributions of errors along radii of Fig. 26 (reference 3), center 0.150, 0.680.

0° filter	90° filter	Number of deviations: Less than 0.675σ	More than 0.675σ	Equal to 0.675σ
86	14	22	24	4
86	14	25	24	1
85	24	25	25	0
30	36	24	21	5
71	24	24	23	3
71	24	27	22	1
78	14	16	26	8
78	14	22	28	0
44	45	19	31	0
44	45	26	24	0
72	15	30	20	0
72	15	23	27	0
54	16	18	28	4
46	35	26	20	4
	Totals	327	343	

and equidistant from that fixed line. If the *constant width* of the strip thus formed is infinitesimal, then the frequency of the matches, as implied by the law (1′), must be of the form

$$f(r)=c\exp(-\varphi^2r^2),$$

where the coefficient c, independent of r, has to be determined from the condition $2\int_0^\infty f(r)dr=1$.* This gives $c=\varphi/\sqrt{\pi}$, so that

$$f(r)=\frac{\varphi}{\sqrt{\pi}}\exp(-\varphi^2r^2),$$

where $\varphi=\varphi(\tilde{\omega})$ is as in (4). For the standard deviation σ, we have

$$\sigma^2=2\int_0^\infty r^2f(r)dr=\frac{2}{\varphi^2\sqrt{\pi}}\int_0^\infty u^2\exp(-u^2)du=\frac{1}{2\varphi^2}.$$

Thus, if we accept the two-dimensional probability function (1) and the stated interpretation of guidance, the color matches in the direction contemplated should have a *frequency distribution*

$$f(r)=\frac{1}{\sigma(2\pi)^{\frac{1}{2}}}\exp(-r^2/2\sigma^2), \qquad (5)$$

that is, a normal distribution, with a *standard deviation* σ determined by

$$\frac{1}{\sigma^2}=\frac{1}{1-\rho^2}\left\{\frac{\cos^2\tilde{\omega}}{\sigma_1^2}-\frac{\rho\sin 2\tilde{\omega}}{\sigma_1\sigma_2}+\frac{\sin^2\tilde{\omega}}{\sigma_2^2}\right\}. \qquad (6)$$

The coefficients σ_1, σ_2, ρ are some characteristics of the color center only, independent of the direction of the line issuing from it, or of the angle $\tilde{\omega}$. Therefore, the points $U=\sigma\cos\tilde{\omega}$, $V=\sigma\sin\tilde{\omega}$, representing the standard deviations in various directions, should lie on the conic

$$\frac{1}{\sigma_1^2}U^2+\frac{1}{\sigma_2^2}V^2-\frac{2\rho}{\sigma_1\sigma_2}UV=1-\rho^2, \qquad (6a)$$

which, in view of $\rho^2<1$, is an *ellipse*.

The latter conclusion is borne out by the experimental results,[3] since the loci of standard-deviation points surrounding the twenty-five color centers,[4] are approximated closely enough by ellipses.

In order to test the former conclusion, claiming a normal distribution (5) of matches along every diameter of the ellipse, we notice that the probability of a match, on any such line, falling within a distance r from the center is

$$p(r)=2\int_0^r f(r)dr=\frac{2}{\sqrt{\pi}}\int_0^{r/\sigma\sqrt{2}}\exp(-u^2)du=\Phi\left(\frac{r}{\sigma\sqrt{2}}\right),$$

where Φ is the error function. E.g., for $r=0.675\sigma$

TABLE II. Distributions of errors along radii of Fig. 44 (reference 3).

0° filter	90° filter	Number of deviations: Less than 0.675σ	More than 0.675σ	Equal to 0.675σ
106	104	21	26	3
119	73	15	35	0
25	81	22	23	5
79	38	18	32	0
79	38	18	27	5
79	38	14	36	0
80	43	19	25	6
42	52	19	31	0
92	53	25	25	0
65	70	14	36	0
37	120	14	21	15
	Totals	199	317	

* Since $f(r)dr$ stands for the probability of a guided color match, r to $r+dr$ in either sense, $\tilde{\omega}$ or $\tilde{\omega}+\pi$.

[4] Reference 3, Figs. 23 to 47.

= "probable error" of the matches,

$$p=\Phi(0.477)=\tfrac{1}{2}.$$

The number of matches, in any direction, with $r<0.675\sigma$ should be equal to the number of matches with $r>0.675\sigma$. Briefly, there should be as many matches *in* as *out*. Of the fifty matches made on every line, some did not differ perceptibly from $r=0.675\sigma$, while the remaining ones were distinctly "in" or "out." Table I shows the numbers of observations which had deviations less than, more than, and approximately equal to 0.675σ for each series indicated in Fig. 26 of reference 3. (Note that σ is identical with Δs shown in Table III of reference 3.)

In Table I the total number of deviations definitely greater than the corresponding "probable errors" is only 2 percent in excess of the number less. The least normal case is analyzed in Table II, for the ellipse centered at 0.441, 0.198 (shown in Fig. 44 of reference 3). As in Table I, the data are given in the same order as in Table III of reference 3.

In Table II, representing the greatest departure from a normal distribution, the total number of deviations greater than the corresponding "probable errors" is 59 percent in excess of the number less. The totals for the numbers of deviations greater and less than 0.675σ for all ellipses are shown in Table III.

Without much ado we may say that, on the whole, the observed *radial* distribution of color matchings does not clash with the frequency formula (5) derived from the two-dimensional normal law (1).

Returning to their transversal distribution around a center, as expressed by formula (6), we compute σ_1, σ_2, and ρ from Eq. (3), and the distribution function given in (1′). In this manner,

$$\sigma_1^2=2C\int_{\tilde\omega=0}^{\pi}\int_{r=0}^{\infty}r^2\cos^2\tilde\omega\exp(-r^2/2\sigma^2)rdrd\tilde\omega$$

$$=4C\int_0^{\pi}U^2\sigma^2d\tilde\omega,$$

where

$$U=\sigma\cos\tilde\omega.$$

As indicated previously, C is evaluated by use of the relation

$$2C\int_0^{\pi}\int_0^{\infty}\exp(-r^2/2\sigma^2)rdrd\tilde\omega=1,$$

from which

$$C=1/2\pi\langle\sigma^2\rangle_{Av}. \tag{3a}$$

Therefore,

$$\sigma_1^2=2\langle U^2\sigma^2\rangle_{Av}/\langle\sigma^2\rangle_{Av}. \tag{3b}$$

Similarly,

$$\sigma_2^2=2\langle V^2\sigma^2\rangle_{Av}/\langle\sigma^2\rangle_{Av}, \tag{3c}$$

where

$$V=\sigma\sin\tilde\omega.$$

Finally,

$$\langle uv\rangle_{Av}=2\langle UV\sigma^2\rangle_{Av}/\langle\sigma^2\rangle_{Av}, \tag{3d}$$

and

$$\rho=\langle uv\rangle_{Av}/\sigma_1\sigma_2. \tag{3e}$$

All of the means indicated in the right-hand terms of (3a) to (3d) are simple averages of the products as functions of $\tilde\omega$. The most probable values of the constants σ_1, σ_2, and ρ are determined by substitution of the experimental values for $\sigma(\tilde\omega)$ in (3a) to (3d). Experimental data are not available for equal intervals of $\tilde\omega$, but if σ is known for n discrete values of $\tilde\omega$, say, $\tilde\omega_1,\cdots\tilde\omega_k\cdots\tilde\omega_n$ between -90 degrees and 90 de-

TABLE III. Summary of analysis of deviations greater and less than "probable error."

Figure (reference 3)	Center x	Center y	Number of deviations Less than 0.675σ	Number of deviations More than 0.675σ	Ratio Number more / Number less
23	0.160	0.057	156	147	0.94
24	0.187	0.118	218	208	0.95
25	0.253	0.125	121	174	1.44
26	0.150	0.680	327	343	1.02
27	0.131	0.521	196	240	1.22
28	0.212	0.550	214	316	1.48
29	0.258	0.450	208	303	1.46
30	0.152	0.365	139	156	1.12
31	0.280	0.385	139	152	1.09
32	0.380	0.498	252	267	1.06
33	0.160	0.200	141	149	1.06
34	0.228	0.250	128	138	1.08
35	0.305	0.323	128	182	1.42
36	0.385	0.393	178	188	1.06
37	0.472	0.399	227	240	1.06
38	0.527	0.350	84	114	1.36
39	0.475	0.300	103	121	1.18
40	0.510	0.236	169	180	1.06
41	0.596	0.283	131	149	1.14
42	0.344	0.284	150	169	1.13
43	0.390	0.237	208	263	1.27
44	0.441	0.198	199	317	1.59
45	0.278	0.223	278	303	1.09
46	0.300	0.163	165	190	1.15
47	0.365	0.153	226	258	1.14
	Totals		4608	5267	1.14

TABLE IV. Computation of constants of normal distribution law.

k	t	$\tilde{\omega}$	$(\tilde{\omega}_{k+1}-\tilde{\omega}_{k-1})=\Delta\tilde{\omega}_k$	$\sigma^2\cdot\Delta\tilde{\omega}_k$	$\sigma^4\cos^2\tilde{\omega}_k\cdot\Delta\tilde{\omega}_k$	$\sigma^4\sin^2\tilde{\omega}_k\cdot\Delta\tilde{\omega}_k$	$\sigma^4\sin\tilde{\omega}_k\cos\tilde{\omega}_k\cdot\Delta\tilde{\omega}_k$
0	7.35	−97.7°					
1	−1.70	−59.5	76.4	69.0	15.9	46.4	−27.2
2	−0.39	−21.3	56.1	72.9	81.9	12.2	−31.7
3*	−0.06	−3.4	35.3	65.2	124.6	0.5	−7.4
4	+0.25	14.0	33.5	120.9	409.4	25.6	+102.5
5	+0.58	30.1	31.6	230.4	1251.0	425.7	729.7
6	1.02	45.6	26.9	156.2	446.1	462.3	454.2
7	1.54	57.0	27.3	84.6	78.1	182.8	119.3
8	3.24	72.9	25.3	38.3	5.0	53.3	16.2
9	7.35	82.3	47.6	55.5	1.3	63.6	8.9
10	−1.70	120.5					
		Totals	360.0	893.0	2413.3	1272.4	1364.5

$\sigma_1^2=2(2413.3)/(893.0)=5.40,\quad \sigma_1=2.32$
$\sigma_2^2=2(1272.4)/(893.0)=2.85,\quad \sigma_2=1.69$
$\langle uv\rangle_{Av}=2(1364.5)/(893.0)=3.06,\quad \rho=3.06/(2.32)(1.69)=0.780.$

* Three sets of fifty observations each. Average of the three values of σ^2 is used in Tables IV and V.

TABLE V. Comparison of calculated and experimental values of σ.

t	−1.70	−0.39	−0.06*	0.25	0.58	1.02	1.54	3.24	7.35
Calculated	0.90	1.07	1.37	1.97	2.69	2.35	1.81	1.32	1.15
Experimental	0.95	1.14	1.36	1.90	2.70	2.41	1.76	1.23	1.08
Difference	−0.05	−0.07	+0.01	+0.07	−0.01	−0.06	+0.05	+0.09	+0.07

* See note to Table IV.

grees, then, for example,

$$\langle\sigma^2\rangle_{Av}=\frac{1}{360}\sum_1^n\sigma_k^2\cdot\Delta\tilde{\omega}_k, \qquad (3f)$$

where

$$\Delta\tilde{\omega}_k=\tilde{\omega}_{k+1}-\tilde{\omega}_{k-1},\quad \tilde{\omega}_0=\tilde{\omega}_n-180°$$

and

$$\tilde{\omega}_{n+1}=\tilde{\omega}_1+180°.$$

Likewise,

$$\langle U^2\sigma^2\rangle_{Av}=\frac{1}{360}\sum_1^n\sigma_k^4\cos^2\tilde{\omega}_k\cdot\Delta\tilde{\omega}_k, \quad \cdots \quad (3g)$$

$$\langle V^2\sigma^2\rangle_{Av}=\frac{1}{360}\sum_1^n\sigma_k^4\sin^2\tilde{\omega}_k\cdot\Delta\tilde{\omega}_k, \quad \cdots \quad (3h)$$

and

$$\langle UV\sigma^2\rangle_{Av}=\frac{1}{360}\sum_1^n\sigma_k^4\sin\tilde{\omega}_k\cos\tilde{\omega}_k\cdot\Delta\tilde{\omega}_k. \quad \cdots \quad (3i)$$

An example of such a computation is given in Table IV for the color center (0.441, 0.198) shown in Fig. 44 of reference 3. For the sake of brevity, the factor 10^{-3} will be omitted in the numerical values of σ. In this table, t signifies the tangent of the angle $\tilde{\omega}$, identical with the slope dy/dx given in Table III of reference 3.

The standard deviation σ, corresponding to any direction, and the directions and values of the maximum and minimum values of σ, can be computed by use of the constants just determined. If we substitute $t=\tan\tilde{\omega}$, then formula (6) becomes

$$(1+t^2)/\sigma^2=\left[\frac{1}{\sigma_1^2}-\frac{2\rho t}{\sigma_1\sigma_2}+\frac{t^2}{\sigma_2^2}\right]\Big/(1-\rho^2), \qquad (6.1)$$

or briefly,

$$\frac{1}{\sigma^2}=(At^2+Bt+C)/(1+t^2). \qquad (6.1)$$

The extreme values of σ are attained when

$$(At^2+Bt+C)t/(1+t^2)=At+B/2$$

or

$$t^2+2(C-A)t/B=1,$$

whence the two mutually perpendicular principal directions or axes of the ellipse,

$$t_m=[A-C\pm\{(A-C)^2+B^2\}^{\frac{1}{2}}]/B, \qquad (6.2)$$

and the corresponding extreme σ values,

$$\frac{1}{\sigma^2}=A+B/2t_m, \qquad (6.3)$$

TABLE VI. Computed values of constants of ellipses.

Color center x	y	$\sigma_1\times10^3$	$\sigma_2\times10^3$	ρ	$A=g_{22}\times10^6$	$B=2g_{12}\times10^6$	$C=g_{11}\times10^6$	$a\times10^3$	$b\times10^3$	θ
0.160	0.057	0.51	0.84	0.756	3.28	−8.15	8.86	0.94	0.30	62.3°
0.187	0.118	0.74	2.23	0.786	0.526	−2.48	4.74	2.31	0.44	74.8
0.253	0.125	1.47	2.03	0.915	1.49	−3.76	2.83	2.49	0.49	54.8
0.150	0.680	2.96	8.86	−0.643	0.0217	0.0836	0.195	9.09	2.21	102.9
0.131	0.521	2.55	4.42	−0.498	0.0679	0.118	0.204	4.67	2.10	110.5
0.212	0.550	2.46	5.55	−0.331	0.0364	0.0543	0.185	5.63	2.30	100.0
0.258	0.450	2.09	4.53	0.045	0.0487	−0.0096	0.230	4.54	2.08	88.5
0.152	0.365	2.20	3.62	−0.463	0.0972	+0.148	0.262	3.81	1.86	111.0
0.280	0.385	1.81	4.12	0.550	0.0843	−0.212	0.439	4.26	1.46	74.6
0.380	0.498	1.93	3.99	0.690	0.120	−0.342	0.512	4.23	1.32	69.4
0.160	0.200	0.95	2.07	−0.160	0.240	0.168	1.14	2.08	0.94	95.4
0.228	0.250	1.28	2.94	0.736	0.252	−0.853	1.34	3.09	0.82	70.9
0.305	0.323	1.49	2.17	0.847	0.750	−1.85	1.59	2.55	0.68	57.2
0.385	0.393	2.03	3.41	0.618	0.139	−0.289	0.391	3.70	1.48	65.5
0.472	0.399	2.17	2.72	0.703	0.268	−0.471	0.421	3.21	1.30	54.0
0.527	0.350	2.42	1.53	0.478	0.551	−0.335	0.222	2.56	1.27	22.8
0.475	0.300	2.57	1.65	0.738	0.807	−0.764	0.332	2.89	0.99	29.1
0.510	0.236	2.16	1.58	0.577	0.600	−0.507	0.323	2.40	1.15	30.7
0.596	0.283	2.46	1.23	0.311	0.740	−0.228	0.184	2.49	1.15	11.1
0.344	0.284	1.28	2.09	0.579	0.346	−0.651	0.917	2.24	0.97	65.7
0.390	0.237	1.86	1.83	0.720	0.622	−0.878	0.598	2.43	0.98	44.2
0.441	0.198	2.32	1.69	0.780	0.896	−1.02	0.473	2.73	0.90	33.7
0.278	0.223	1.29	2.08	0.844	0.807	−2.19	2.10	2.34	0.61	60.3
0.300	0.163	1.89	2.47	0.919	1.05	−2.53	1.81	3.01	0.60	53.4
0.365	0.153	3.18	2.60	0.900	0.780	−1.15	0.519	4.12	0.90	38.6

where

$$A,\ B,\ C=\left(\frac{1}{{\sigma_2}^2},\ -\frac{2\rho}{\sigma_1\sigma_2},\ \frac{1}{{\sigma_1}^2}\right)\Big/(1-\rho^2).$$

Using the values determined in Table IV, formula (6.1) becomes

$$\sigma^2=(1+t^2)/(0.896t^2-1.016t+0.473).$$

The experimental values of σ are compared with the values computed from the formula in Table V.

The values of σ_1, σ_2, and ρ for the color center 0.441, 0.198, determined in Table IV, can be used in (6.2) to determine the most probable angle of inclination of the major axis of the ellipse in Fig. 44 of reference 3. In this case, $A=0.896$, $B=-1.016$, $C=0.473$, and t_m has the two values 0.667 and -1.50. To the former corresponds the maximum, and to the latter, the minimum of σ. Their values are given by (6.3), and the inclination of the major axis of the ellipse to the x axis is $\theta=\arctan 0.667$. Thus, the semi-axes and the orientation of the most probable ellipses fitting the data are,

$$a=2.73,\quad b=0.90,\quad \theta=33.7°.$$

These values are in almost perfect agreement with the ellipse shown in Fig. 44 of reference 3, which is merely the ellipse which in MacAdam's judgment most closely fitted the indicated points. The method of computation shown in Table IV is based on the two-dimensional normal distribution law. By eliminating dependence on personal judgment, this method confirms and strengthens the conclusions previously based on the ellipses.

The computed values of σ_1, σ_2, ρ, A, B, C, a, b, and θ for all twenty-five ellipses are shown in Table VI. The values for the first and thirteenth ellipses listed are less reasonable than those for the ellipses shown in Figs. 23 and 35 of reference 3 because of the paucity of data midway between the axes of the ellipses. In such cases, the eccentricities of the ellipses are exaggerated by the computations based on formulae (3g) to (3i).

Summing up, we may say that the bulk of MacAdam's experimental results is substantially accounted for by the assumption of *normal frequency distribution* of the two attributes $x-x_0$, $y-y_0$ in matching any center x_0, y_0 of the manifold of equiluminous colors or, equivalently, that Cotes's centroid postulate holds around every point of this twofold. The particular normal frequency function belonging to a color (x_0, y_0) is characterized by the three coefficients σ_1, σ_2, and ρ whose values vary from color to color. The

correlation coefficient ρ, which may be positive or negative, has the essential property $\rho^2<1$. These three coefficients determine completely the size, shape, and orientation of the standard-ellipse surrounding each color.

The validity of the normal law of frequency for the color domain, thus established, enables us at once to construct its element of distance or its "line element" ds^2 as a definite quadratic differential form of the coordinates x, y. If $P(x, y)$ and $Q(x+dx, y+dy)$ are any two points of the twofold, their charted distance apart is $\mathbf{PQ}=(dx^2+dy^2)^{\frac{1}{2}}$. As explained before, their true distance ds is defined by

$$ds=\mathbf{PQ}/\sigma \quad \text{or} \quad ds^2=(dx^2+dy^2)/\sigma^2,$$

where σ is the standard deviation of color matching in the direction $\mathbf{PQ}$. Hence, by formula (6), and since $\mathbf{PQ}\cdot\cos\tilde{\omega}=dx$, $\mathbf{PQ}\cdot\sin\tilde{\omega}=dy$,

$$ds^2=\left(\frac{dx^2}{\sigma_1^2}-\frac{2\rho}{\sigma_1\sigma_2}dxdy+\frac{dy^2}{\sigma_2^2}\right)\Big/(1-\rho^2), \quad (7)$$

or

$$ds^2=g_{11}dx^2+2g_{12}dxdy+g_{22}dy^2, \quad (7a)$$

where

$$g_{11}, g_{12}, g_{22}=\left(\frac{1}{\sigma_1^2}, -\frac{\rho}{\sigma_1\sigma_2}, \frac{1}{\sigma_2^2}\right)\Big/(1-\rho^2).$$

Since $\rho^2<1$, we have always $g_{11}g_{22}-g_{12}^2>0$, so that (7a) is a definite, positive quadratic form. The components g_{ik} of the metrical tensor of the color twofold are here expressed by the partial standard deviations σ_1, σ_2 and the correlation coefficient ρ which depend, of course, on the choice of the system of stimuli-coordinates x, y. If we pass from these to any other coordinates x', y' by any transformation whatever, the coefficients σ_1', σ_2', ρ' and, therefore, also the corresponding tensor components g_{ik}' will differ from the previous ones, but the true distance of any two colors, $ds=\mathbf{PQ}/\sigma'$, will remain unchanged. The line element, as defined above, is invariant with respect to any coordinate transformations.

Notice, however, that the normal law (1) will now be expressed by

$$2\log\frac{f_0}{f}=g_{11}(x-x_0)^2+2g_{12}(x-x_0)(y-y_0)+g_{22}(y-y_0)^2;$$

if, therefore, g_{ik} is to be a general tensor, the deviations $x-x_0$, $y-y_0$ must be considered as infinitesimal.* The probability function is thus reduced to $f=f_0\exp(-ds^2/2)$.

Keeping this clause well in mind, we now see that the possibility of treating the domain of equiluminous colors as a geometrical twofold or surface is essentially conditioned by the (at least, approximate) validity of the normal law of frequency of color matchings.

The "normal" law of frequency for two variable attributes (as, indeed, for a single attribute) is by no means the rule, but rather the exception. In biometrics, many cases are known where it does not hold even approximately. It happens to hold, closely enough, for the domain of equiluminous colors, at least for MacAdam's observers, and this circumstance only enables us to represent its metrics by a proper line element, $ds^2=g_{11}dx^2+2g_{12}dxdy+g_{22}dy^2$, or to consider it as what the mathematicians call a *Riemannian twofold*.

For three variable attributes, X, Y, Z, the normal law of frequency, which is again derivable from Cotes's centroid postulate, is of a similar form, namely, $\log f_0/f=$ a homogeneous quadratic function of $X-X_0$, $Y-Y_0$, $Z-Z_0$, containing, in general, three different "correlation" coefficients, between the attributes X and Y, Y and Z, Z and X. Whether this law holds for the complete color domain, with its stimuli-coordinates as the three attributes, we do not know. This question, of fundamental importance, could be decided, perhaps, by a thorough analysis of the wealth of color matchings, with independent adjustments of three coordinates, accumulated by Guild.[5] If the normal law does hold for this domain, then the matchings of a color center in any direction $X:Y:Z$ will show a distribution analogous to (5), with a standard deviation σ

---

* Otherwise, g_{ik} would be but a special tensor and the "true distance" ds of colors would be invariant only with respect to *linear* transformations of the stimuli-coordinates. By "infinitesimal" $x-x_0$, $y-y_0$ are meant intervals within which the coefficients g_{ik} do not change appreciably, and thus, intervals not much greater than the standard deviation σ. It is true that we have, on one or two occasions, integrated the probability function over r up to infinity. But the values of these integrals would remain very nearly the same, if their upper limit $r=\infty$ were replaced by some 3σ or 4σ only. Thus, the semblance of inconsistency is removed.

[5] J. Guild, Phil. Trans. Roy. Soc. **A230**, 149–187 (1931).

varying with the direction in such a manner that the locus of σ points will be an ***ellipsoid*** with the color matched as center and with the size and orientation of the principal axes varying in some intricate way from color to color. If the elementary distance ds of two colors is again defined as their charted separation divided by the corresponding standard deviation σ (or, possibly, some power of σ), then ds^2 will be a definite differential quadratic form of the stimuli-coordinates,

$$ds^2 = g_{11}dX^2 + 2g_{12}dXdY + \cdots + g_{33}dZ^2,$$

with its six tensor components $g_{ik} = g_{ki}$ varying from color to color. (The condition for the general tensor character of g_{ik} will be as before.) In other words, the complete color domain will be a *Riemannian threefold*, as assumed, tentatively, by Helmholtz and his followers. The hitherto unproved admissibility of Helmholtz's assumption is thus reducible to the question of validity of the normal law of frequency for the three-color attributes, which can be answered only by a patient discussion of experimental color matchings.

The point of view adopted here and substantiated, so far, for equiluminous colors is, perhaps, best characterized by the inversion of the formula $f = f_0 \exp(-\frac{1}{2}ds^2)$ found for the probability function or frequency of color matchings. It gives for the line element of the color domain or the squared visual sensitivity to color difference, in terms of the directly observable frequency of matchings,

$$ds^2 = 2 \log f_0/f.*$$

This, then, replaces Fechner's debatable psychophysical law (that sensation is proportional to the logarithm of the stimulus) which has served as a basis in Helmholtz's construction of the metrical color domain.

* Notice, incidentally, that $f_0 = \sqrt{g}/\pi$, where g is the determinant of the metrical tensor, and since (in 2 dimensions) $\sqrt{g} = 1/ab$, $f_0 = 1/\pi ab$, the reciprocal of the charted area of the standard ellipse surrounding the color x_0, y_0.

Ultraviolet Irradiation of Interrupter Gaps in Spectrographic Spark Sources

S. F. SIMPSON
Chevrolet Motor and Axle Division, General Motors Corporation, Buffalo, New York
(Received August 11, 1944)

IN May, 1942 the high voltage spark unit used by the author was subjected to a series of tests to determine the effect of ultraviolet irradiation of the spark gaps at the synchronous interrupter. This spectrographic source is a standard commercial unit of modified Feussner[1] type. It has a variable resistance and a tapped inductance in the primary; and a secondary circuit consisting of a 35,000-volt 2-kva transformer, maximum capacitance of 0.021 μf, tapped inductance, interrupter, safety gap, and sample gap. For these original tests a 6″ ultraviolet lamp was borrowed from Sun-Lamp service and positioned directly in front of the interrupter.

We started the spark unit, adjusted the voltage drop of the primary to 75 volts, and then turned on the ultraviolet lamp. The normally slow starting lamp reached full brilliancy almost immediately, and the effect on the spark unit was quite pronounced. A quieter, more uniform spark tone was very apparent. The flaring and irregularities, ordinarily noticeable because of the stroboscopic illumination of the interrupter wheel, became lower in magnitude and fewer in number. We also observed a primary voltage reduction to 63 volts. We increased the voltage to the original 75 volts by moving the stationary contacts in the direction of the wheel rotation, without losing the improvement in tone and appearance of the spark.

A series of tests was run on precision of analysis with and without the irradiation. Five spectra were taken with each condition on each of several plates over an extended period. Reproductivity, calculated as standard deviation divided by concentration, showed a consistent improvement of 40 percent or greater on all determinations made during the intervals when the ultraviolet was used.

After consideration of these data we decided to purchase a lamp for permanent installation. Practical tests were made to determine the strength of irradiation needed. The original lamp was tested at various positions and from these observations optimum strength was calculated. A 250-watt, 3″ mercury arc, in a fused quartz tube was chosen as ideal for the limited space available. With the special transformer, starting cut-out, and socket required, the cost was less than $50.00. A full year's observation was made on routine use of the lamp, during which various installation positions and other conditions were tested, with the following conclusions. The lamp should have bright, firm, contact clips. It should be reasonably free from vibration, and away from the path of centrifugally thrown dust, oil, metallic oxide, etc., from the interrupter wheel. If a reflector is used it should be short, as open as possible, and plated for longest wear and greatest reflection of the valuable wave-lengths. The only machineable base material tried, which would stand the heat, was transite. Copper fuse clips are used as holders, with the lamp ends approximately 6″ from the gaps in the latest assembly. This distance was chosen by trial tests during which we observed a point of maximum effect, checked as precision of analysis, and a reduction of effect as the lamp was moved from the gaps.

The practical application of this lamp to our spark unit has been of great value to our analyses of aluminum, magnesium, and steel alloy constituents in high concentrations. Errors caused by excitation have been so reduced that this phase of the method is no longer the major source of deviations. Certain other changes have been developed in the spark unit which also contributed to reduction of errors, but the tests mentioned were all made with variation only in the irradiation, and its application remains the greatest single improvement made in our excitation.

The theoretical reasons for the effect created by irradiation of the interrupter are based on well-known and studied principles. Many observers have contributed to the data on electrical discharges, although only a few direct references

[1] O. Feussner, Archiv. f. Eisenhutt. 6, 551 (1932).

are given in our bibliography. A spark is dependent on distance, pressure, and the availability of free ions in the gap space; and the frequently reported variations from the physical laws governing such discharges are most often caused by the latter condition. The great variations in breakdown voltages and the time lags exhibited by excessive impulse voltages are both eliminated by irradiation of the gap with an ionizing radiation.[2] We might then have presumed that our ultraviolet lamp would similarly improve any spark circuit, but found that this was not the case. Tested on a circuit of similar type but different construction[3] the improvement was negligible. Another effect than more uniform ionization was then sought, and a closely related phenomenon was studied.

The photoelectric effect[4] lowers the voltage necessary to jump the closing gaps between the moving and the stationary electrodes, thus advancing spark discharge in the sine wave. This, if not corrected, changes the excitation characteristics and may cause considerable shifting of analytical curves. It is prevented by moving the peripheral positions of the stationary electrodes, or by moving the wheel on its shaft, until the same primary voltage drop is recorded as formerly. Effectively then the current density, resistance, and frequency of the circuit are back to "normal," and only the position of the discharge in relation to the coincidence of the approaching electrodes is altered. The spark starts farther from the stationary electrodes, but being of equal strength and frequency it is of equal duration (compared to the un-ionized discharge) and is more likely to be completed before coincidence of electrodes is reached. Completion of the wave, however, is not necessary to excitation as long as the profile of the wave is uniformly reproduced. Our inability to do this without preliminary ionization is attributed partially to the variations mentioned in the preceding paragraph, but largely to the periodicity of the driving motor. Figure 1 diagrams the possible effects of a momentarily fast or slow motor.

The wheels are shown in sequence, 10° apart, along an axis representing the position in the half-wave where the condenser has a standard charge. The timing is exaggerated to simplify the diagram. *B* and *C* might indicate the beginning and end of a perfect spark, assuming no irradiation and absolute synchronism of the motor. *A* indicates a slow position of the wheel, when the safety gap would probably be discharged. The wheel is originally positioned where these safety gap discharges are few. This setting is indicated by the left end of the lower shaded band. A typical current wave is diagrammed to indicate a unit length. This wave is directly responsible for the emitted light, and any change in its length, shape, or frequency indicates a corresponding change in excitation.[5] A discharge wave must be completed before coincidence of electrodes, as in *C*, or the greatly reduced gap length will alter its critical characteristics. In one typical case a fast wheel *C*, starting its discharge later than the right end of the lower shaded section, would produce lower current densities, higher frequencies, and probably an abnormal wave profile if the gap reached or passed its minimum width during the discharge.

With the irradiation the stationary electrodes must be repositioned (indicated by the filled circles) to prevent an early discharge of low current density. This was done so that previously made analytical curves might still be used. We observed a reduction of safety gap discharges, indicating that the ionizing of the spark gaps was compensating better than in the previous condition for slow wheel positions. This indicates that the upper shaded section of our time chart might safely be started further to the left. The new electrode position indicates that the full discharge can now be completed before coincidence of electrodes, even on fast wheel positions, as in *C*. The longer discharge interval is thus able to overcome the irregular effects caused by the momentary acceleration or deceleration of the motor during its individual cycle.

This theory is further substantiated by the trial of an ultraviolet source at the sample gap, which showed no benefit to analytical precision. The assumption was then made that the chief cause of excitation error in the spark unit used

[2] T. E. Allibone and J. M. Meek, J. Sci. Inst. **21** (Feb. 1944).
[3] H. B. Vincent and R. A. Sawyer, J. Opt. Soc. Am. **32**, 686 (1942).
[4] H. Hertz, Ann. d. Physik **31**, 983 (1887).
[5] H. Kaiser and A. Wallraff, Ann. d. Physik **34**, 297 (1939).

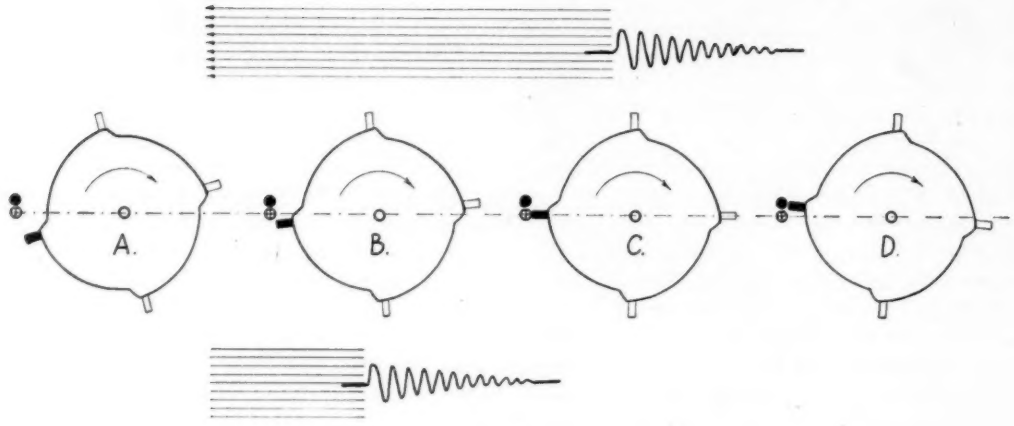


FIG. 1

for these tests was in the lack of perfect synchronism of the small driving motor. Small induction motors are particularly sensitive to phase shifting and periodicity (because of the individual characteristics of the motor armature, the alternator, or the prime mover), and they constantly "hunt" for the synchronous position. This surge or lag may sometimes be minimized by reduction of the field current, by use of shunts, or by use of a heavier motor. Tests of this theory using a heavier motor proved it correct, and provided further improvement toward uniform excitation.

Further improvements have been made in our spark unit upon consideration of later observations. A new interrupter wheel was built using easily adjustable, pencil-type electrodes. Aerodynamic principles were considered in designing its periphery. First, we attempted to eliminate by vacuum the space charge residue from each discharge to prevent its effect on the one following. Second, we sought to provide a pressure wave, timed to encourage each discharge. This applies Paschen's law that discharge voltage gradient is the product of pressure and distance.

Numerous other changes have been made but none has been of greater benefit to excitation than ultraviolet irradiation of the interrupter in this spark unit.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 35, NUMBER 1 JANUARY, 1945

A Modified Photochemical Theory of Vision

PARRY MOON, *Massachusetts Institute of Technology, Cambridge, Massachusetts*

AND

DOMINA EBERLE SPENCER, *Tufts College, Medford, Massachusetts*

(Received September 14, 1944)

1. INTRODUCTION

DESPITE the large amount of excellent research that has been done on vision, the subject remains in an undeveloped state. There are two general methods of eliminating this unsatisfactory condition:

(1) By resorting to further experiment,
(2) By making bold hypotheses, the predictions from which are checked experimentally at salient points.

Needless to say, the two methods are supplementary. But a reliance on pure experiment without theory—which is the traditional approach in visual research—seems to be even less fruitful in the study of vision than in physics. There are several reasons for this fact:

(*a*) Visual phenomena are extremely complex, with a large number of independent variables. The scientific attack consists in fixing all the independent variables but one, which is then changed in a known manner. The complexity of visual problems is such that the complete experimental investigation of any phenomenon has never been attempted. Usually only one of the many variables is changed, and a hope is expressed that the remaining variables will be fixed or at least will not vary so widely as to invalidate the results. Sometimes the hope is realized. In visual-acuity research, for instance, helios is ordinarily the only quantity that is varied. To fix a single curve in this way may require a hundred thousand readings, and to obtain visual acuity as a function of contrast, color, and all the other pertinent variables becomes almost hopeless.

(*b*) Many experiments on vision are difficult or impossible because the eye is part of a living organism.

(*c*) Reproducibility of measurements is limited by the organism itself, so that the most refined experimental technique does not necessarily result in accurate data.

(*d*) The results depend on the physiological and psychological condition of the observer and change from time to time in an unpredictable manner.

In view of these characteristics, it seems that a postulational approach is particularly suitable in the study of vision. Some investigators are suspicious of hypotheses and feel that the "brute-strength" method of obtaining data for all possible conditions is the only safe procedure. Without some kind of scaffolding afforded by theory, however, it is difficult to know what experiments will be most valuable. We have used the postulational method in several papers[1] and will now apply it to the photochemical theory of vision.

The photochemical theory of vision has been developed by Pütter,[2] Lasareff,[3] and Hecht.[4] A weakness has been that ordinarily a new set of assumptions has been found necessary in applying the theory to each kind of visual test. Such a procedure is unfortunate since it tends to destroy that important function of any theory—the prediction of new results and the suggestion of new avenues of experimental approach. It is found that a modification of the Hecht theory allows the quantitative prediction of a large group of visual phenomena on the basis of a single set of postulates.

The purpose of the present paper is to determine how much of visual theory can be correlated by means of a few basic postulates. It is found that the absolute threshold, the instantaneous

---

[1] P. Moon and D. E. Spencer, J. Opt. Soc. Am. **33**, 444 (1943); **34**, 46 (1944); **34**, 93 (1944); **34**, 234 (1944). D. E. Spencer, J. Opt. Soc. Am. **33**, 10 (1943).

[2] A. Pütter, Arch. ges. Physiol. **171**, 201 (1918).

[3] P. Lasareff, Pflüg. Arch. **213**, 256 (1926).

[4] S. Hecht, J. Gen. Physiol. **18**, 767 (1935); **4**, 113 (1921); **10**, 781 (1927). "The nature of the photoreceptor process," in *Handbook of General Experimental Psychology* (Worcester, Massachusetts, 1934), C. Murchison, Editor.

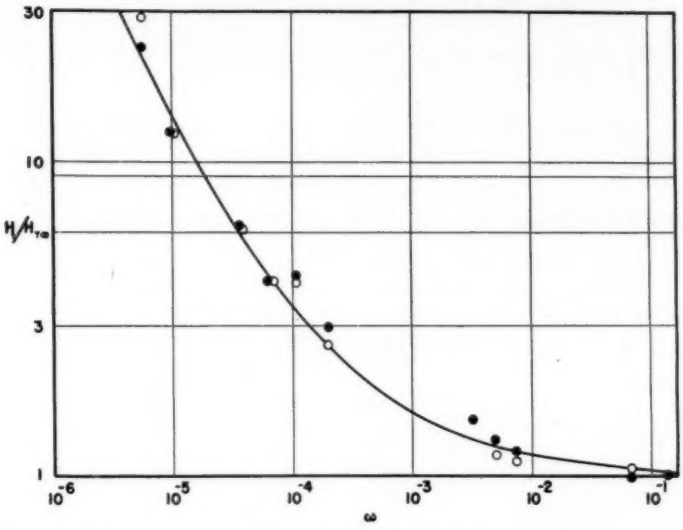


FIG. 1. Effect of area. Cone vision, data of Steinhardt (reference 18). ● $H_A=21$ blondels; ○ $H_A=89$; — Calculated, Eq. (26).

threshold, and numerous phenomena in dark adaptation and light adaptation are satisfactorily handled on a simple postulational foundation. It seems likely that a few additional postulates will extend the range of the theory to include also minimum visual angle and other important visual data.

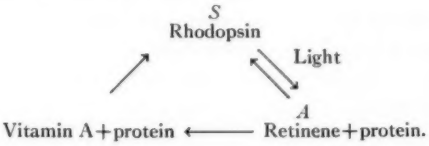
2. POSTULATES

The supposition that the initial phenomenon in vision is photochemical has received so much experimental support that today it can hardly be doubted. Hecht[4] assumes that associated with each receptor element in the retina is a sensitive substance S which is acted on by light to form (among others) a substance A. There are two principal reactions:

(a) A *photochemical* reaction in which S is decomposed and A is formed,

(b) A *chemical* reaction in which S is produced.

Recent researches, particularly those of Wald,[5] have emphasized the fact that the actual phenomena are more complicated than those assumed in the photochemical theory of vision. In the rods of the human retina, the substance S is a rose-colored liquid called *rhodopsin*, and a decomposition product A is the yellow *retinene*. The schematic relationship for the rods, as given by Wald,[5] is as follows:



The two reactions between S and A are those listed as (a) and (b) above.

It is known that the reaction by which retinene changes to vitamin A and the reaction by which vitamin A forms new rhodopsin are both very slow compared with the reactions (a) and (b). Thus it seems permissible to neglect the reactions on the left and bottom of the diagram and confine the theory to (a) and (b). In the cones, similar reactions are supposed to take place but the chemicals are different. The exact structures of the chemical substances and the nature of the chemical reactions are not known for either rods or cones. But, fortunately, a mathematical theory of vision can be developed without recourse to these chemical details.

The mathematical formulation requires equations to represent the reaction velocities of (a) and (b). Additional assumptions must then be

[5] G. Wald, Biol. Symposia **7**, 43 (1942); *Medical Physics* (Year Book Publishers, Chicago, 1944), Otto Glasser, Editor, p. 1658.

made regarding the frequency of impulses in the optic nerve fibers and the relation of this frequency to the visual threshold. The effect of the area of the stimulus must be considered also. These considerations will now be introduced by means of postulates.

Postulate I. The velocity of the photochemical reaction (a) is given by the equation,

$$\left(\frac{dx}{dt}\right)_1 = k_1 H(t)[\exp(-ax) - \exp(-a)] \cdot \zeta(\delta, 2; t), \quad (1)$$

where x = number of molecules of sensitive substance S in one rod or one cone, expressed as a fraction of the maximum possible number; t = time (sec.); $H(t)$ = helios (blondel), at time t, of the object that is imaged on the rod or cone in question; δ = pupil diameter (mm); $\zeta(\delta, 2; t)$ = Stiles factor[6] relating natural pupil at time t with artificial pupil of 2-mm diameter; a, k_1 = constants determined for a fixed pupil diameter of 2 mm.

TABLE II. Visibility of stars.

Authority	Obs.	D_T at eye (lm/m²)	Star magnitude	Distance for candle (km)
Actual Stars				
Ives[a]	Many	85×10^{-10}	6.0	10.8
Artificial Stars				
Reeves[b]	P.R.	20.0×10^{-10}	7.6	22.4
Reeves[b]	Av. of 3	24.4	7.4	20.2
Russell[c]	H.N.R.	8.5	8.6	34.3
Buisson[d]	H.B.	13.8	8.0	27.0
Bouma[e]	——	10.0	8.4	31.6

[a] See reference 24 of text.
[b] P. Reeves, Astrophys. J. **46**, 167 (1917).
[c] H. N. Russell, Astrophys. J. **45**, 60 (1917).
[d] H. Buisson, Astrophys. J. **46**, 296 (1917).
[e] P. J. Bouma, Philips Tech. Rev. **5**, 296 (1940). S. Hecht, J. Opt. Soc. Am. **32**, 42 (1942). Hecht, Shlaer, and Pirenne, J. Gen. Physiol. **25**, 819 (1942). Hecht, Schlaer, and Pirenne, Science **93**, 585 (1941).

Postulate II. The velocity of the chemical reaction (b) is

$$\left(\frac{dx}{dt}\right)_2 = -\left\{k_2 - k_4 x^3 + A \exp\left[-c\int_{-\infty}^{t} \frac{ak_1 H(s)\zeta(\delta, 2; s)\exp[-b(t-s)]}{1+ak_1 H(s)\zeta(\delta, 2; s)s}ds\right]\right\}x. \quad (2)$$

In Eq. (2), s is a variable of integration and corresponds to a value of time at which the helios is $H(s)$. The quantities k_2, k_4, b, c, A are constants determined with a constant pupil diameter of 2 mm.

Postulate III. The resultant reaction velocity is the sum of Eqs. (2) and (3), or

$$\frac{dx}{dt} = \left(\frac{dx}{dt}\right)_1 + \left(\frac{dx}{dt}\right)_2. \quad (3)$$

This is the differential equation whose solution gives the amount of decomposition product as a function of time under all conditions.

Postulate IV. The action of light produces electrical impulses in the nerve fiber associated with the receptor element. The number of impulses per second is

$$f = k_3\left(\frac{dx}{dt}\right)_1, \quad (4)$$

where k_3 is a constant.

Postulate V. There is a minimum frequency f_T below which no sensation occurs. If the number of excited nerve endings is very large, the threshold is reached when

$$f_T = \text{const.} = f_\infty. \quad (5)$$

If the object is reduced in size so that a lesser number of nerve endings is excited, a higher threshold frequency will be necessary:

$$f_T = \frac{f_\infty}{\omega}(A_3 + \omega^{\frac{1}{2}})^2, \quad (6)$$

where f_∞ = threshold frequency for very large ω, ω = solid angle (steradians) subtended by the source, A_3 = constant.

TABLE I. Effect of size of stimulus. Rod vision, circular disk field, data of Wald,* calculated values from

$$H_T = \frac{3.60\times10^{-6}}{\omega}[0.0469+\omega^{\frac{1}{2}}]^2.$$

ω (steradians)	log H_T Obs., 15°	Obs., 25°	Calc.
0.203×10^{-3}	$\bar{5}.82$	$\bar{5}.99$	$\bar{5}.820$
0.848	$\bar{5}.24$	$\bar{5}.43$	$\bar{5}.390$
1.91	$\bar{5}.14$	$\bar{5}.16$	$\bar{5}.189$
3.52	$\bar{5}.06$	$\bar{5}.03$	$\bar{5}.066$
5.87	$\bar{6}.97$	$\bar{6}.98$	$\bar{6}.971$

* See reference 17 of text.

[6] P. Moon and D. E. Spencer, J. Opt. Soc. Am. **34**, 319 (1944); **34**, 744 (1944).

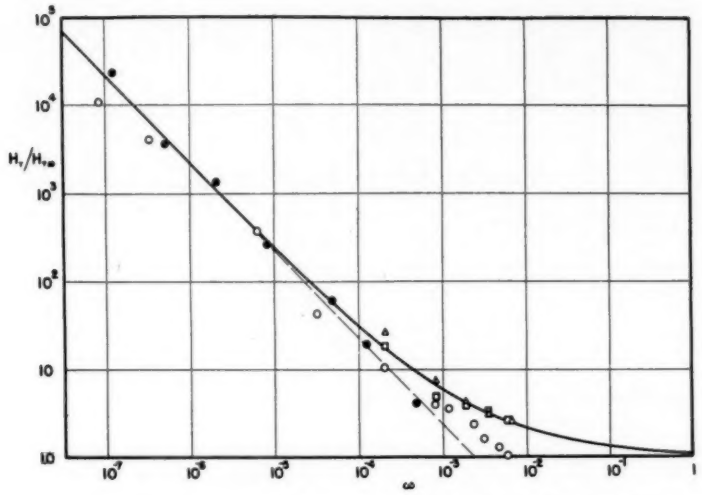


FIG. 2. Effect of area. Rod vision. ○ Reeves (reference 19); ● Houstoun (reference 20); □ Wald (reference 17), fixation 15° from fovea; △ Wald (reference 17), 25° from fovea; — Calculated, Eq. (27); -- Calculated, Eq. (21).

Additional discussion of these postulates will now be given. Postulate I states that the velocity of the photochemical reaction is directly proportional to the amount of light. This relation is in accordance with the Einstein photochemical equivalence principle. It has also been checked experimentally for rhodopsin in solution.[5] Hecht[4] used the simple equation,

$$\left(\frac{dx}{dt}\right)_1 = k_1 H(1-x),$$

in place of our Eq. (1). We have shown, however, that his equation does not fit the modern data on dark adaptation.[7] The more complicated function of x in Eq. (1) allows a quantitative agreement between theory and experiment.

The remaining function, $\zeta(\delta, 2; t)$ in Eq. (1), is introduced to compensate for change in pupil size. Let the effective luminous flux through the pupil, from a large source of helios H, be denoted by F_e. Then, in accordance with the Einstein principle,

$$\left(\frac{dx}{dt}\right)_1 = F_e \cdot \Phi_1(x). \tag{7}$$

With a 2-mm pupil, this flux is directly proportional to the helios H, or

$$F_{e2} = k_5 H.$$

But, according to the theory of the Stiles-Crawford effect*

$$\left.\begin{aligned} F_{e2} &= 2\pi k H \int_0^1 \eta(x)x\,dx \\ F_{e\delta} &= 2\pi k H \int_0^{\delta/2} \eta(x)x\,dx \end{aligned}\right\},$$

and

$$F_{e\delta}/F_{e2} = \frac{\int_0^{\delta/2} \eta(x)x\,dx}{\int_0^1 \eta(x)x\,dx} = \zeta(\delta, 2).$$

Therefore,

$$F_{e\delta} = F_{e2}\zeta(\delta, 2) = k_5 H \zeta(\delta, 2).$$

Substitution in Eq. (7) gives

$$\left(\frac{dx}{dt}\right)_1 = k_5 H \zeta(\delta, 2) \cdot \Phi_1(x), \tag{8}$$

which is of the form of Eq. (1).

When a 2-mm fixed pupil is employed, all the ζ's of Postulates I and II are equal to unity and may be omitted. With a natural pupil, or an artificial pupil with $\delta \neq 2$, each H that occurs in

[7] P. Moon and D. E. Spencer, "Visual dark adaptation; a mathematical formulation," to be published.

* Reference 6, Eqs. (11a) and (11b).

the equations must be multiplied by a Stiles factor ζ. In Eq. (1), the $\zeta(\delta, 2; t)$ is the Stiles factor at time t. It may be calculated by means of the equations in Section 6 of this paper. In Eq. (2), the Stiles factor $\zeta(\delta, 2; s)$ refers to time s.

In place of Postulate II, Hecht used the equation,

$$\left(\frac{dx}{dt}\right)_2 = -k_2 x^2.$$

Winsor and Clark[7, 8] proved that this equation or any similar equation cannot accord with the results of modern dark-adaptation research. Equation (2) differs from all previous formulations by making the reaction velocity depend on the previous light-adaptation history of the eye. Further discussion of this equation is given in another paper.[7]

Postulate III states that only two reactions will be considered: the photochemical reaction (a) and the chemical reaction (b). It is well known that the stimulation of nerve occurs, not continuously but in discrete impulses. Oscillograms of impulses in single nerve fibers[9] show that the magnitude is constant but that the frequency changes with the amount of light. In Postulate IV we assume the simplest possible relation between frequency of impulses and their cause, namely, a linear relation between f and $(dx/dt)_1$. In fact, one may assume that each time a molecule of photosensitive substance is dissociated, it excites a nerve ending and initiates an impulse.[7]

In Postulate V, the effect of number of receptors is considered. The background is assumed to have zero helios. A spot of helios H and solid angle ω is observed and its helios is increased until the spot is barely visible. If the spot is very large so that a great number of receptors are stimulated simultaneously, experiment shows that a minimum helios is required. That is, the threshold frequency f_T has a very small value which is represented by f_∞. This value is independent of ω over a considerable range, but if ω is decreased sufficiently the total visual effect of the spot is reduced unless the frequency of the individual receptors is increased. For the extreme case of very small ω, the frequency varies inversely as the solid angle. The simplest expression to include both of these special cases appears to be of the form of Eq. (6). This subject is discussed in Section 5.

TABLE III. Absolute threshold (rod vision, natural pupil).

Authority	Obs.	Pupil	Fixation	$H_{T\infty}$ (blondel)
Hecht, Haig, and Chase[a]	3	2 mm	30° from fovea	9.5×10^{-6}
Haig[b]	E.M.H.	2 mm	7° from fovea	8.5×10^{-6}
Wald[c]	G.W.	—	15° and 25°	3.6×10^{-6}
Wald & Clark[d]	A-B.C.	Nat.	9.8° from fovea	10.4×10^{-6}
Nutting[e]	P.G.N.	Nat.	None	14.0×10^{-6}
Cobb[f]	15	Nat.	—	(3 to 30) 10^{-6}
Hecht and Mandlebaum[g]				
Fig. 5	—	Nat.	7° from fovea	6.3×10^{-6}
Fig. 6	J.M.	Nat.	7°	5.6×10^{-6}
Fig. 6	L.W.	Nat.	7°	5.6×10^{-6}
Hecht and Mandlebaum	F.W.	Nat.	7°	14×10^{-6}
Bouma[h]	—	Nat.	None	10×10^{-6}

[a] See reference 29 of text.
[b] See reference 30 of text.
[c] See reference 17 of text.
[d] See reference 27 of text.
[e] P. G. Nutting, Trans. I.E.S. **11**, 1 (1916).
[f] P. W. Cobb, Psychol. Rev. **26**, 428 (1919). P. W. Cobb, Arch. Ophth. **48**, 492 (1919).
[g] See reference 25 of text.
[h] P. J. Bouma, Philips Tech. Rev. **5**, 296 (1940). S. Hecht, J. Opt. Soc. Am. **32**, 42 (1942). Hecht, Shlaer, and Pirenne, J. Gen. Physiol. **25**, 819 (1942). Hecht, Shlaer, and Pirenne, Science **93**, 535 (1941).

TABLE IV. Absolute threshold (cone vision, natural pupil).

Authority	Obs.	Pupil	Fixation	$H_{T\infty}$ (blondel)
Hecht, Haig, and Chase[a]	3	2 mm	30° from fovea	72.3×10^{-3}
Hecht[b]	15	Nat.	Fovea	3.0×10^{-3}
Hecht and Mandelbaum[c]				
Fig. 5	1	Nat.	7° from fovea	2.1×10^{-3}
Fig. 6	J.M.	Nat.	7° from fovea	3.2×10^{-3}
Fig. 6	L.W.	Nat.	7° from fovea	1.8×10^{-3}
	F.W.	Nat.	7° from fovea	3.2×10^{-3}

[a] See reference 29 of text.
[b] S. Hecht, J. Gen. Physiol. **4**, 113 (1921).
[c] See reference 25 of text.

3. SOLUTION OF THE DIFFERENTIAL EQUATION

According to the Postulates I, II, and III, the differential equation of vision is

$$\frac{dx}{dt} = k_1 H(t)[\exp(-ax) - \exp(-a)]\zeta(\delta, 2; t) - \left\{k_2 - k_4 x^3 + A \exp\left[-c\int_{-\infty}^{t} \frac{ak_1 H(s)\zeta(\delta, 2; s)\exp[-b(t-s)]}{1 + ak_1 H(s)\zeta(\delta, 2; s)s}ds\right]\right\}x. \quad (9)$$

[8] C. P. Winsor and A-B. Clark, Proc. Nat. Acad. Sci. **22**, 400 (1936).
[9] H. K. Hartline and C. H. Graham, J. Cell. and Comp. Physiol. **1**, 277 (1932). C. H. Graham, *Handbook of General Experimental Psychology* (Clark University Press, Worcester, 1934), p. 829.

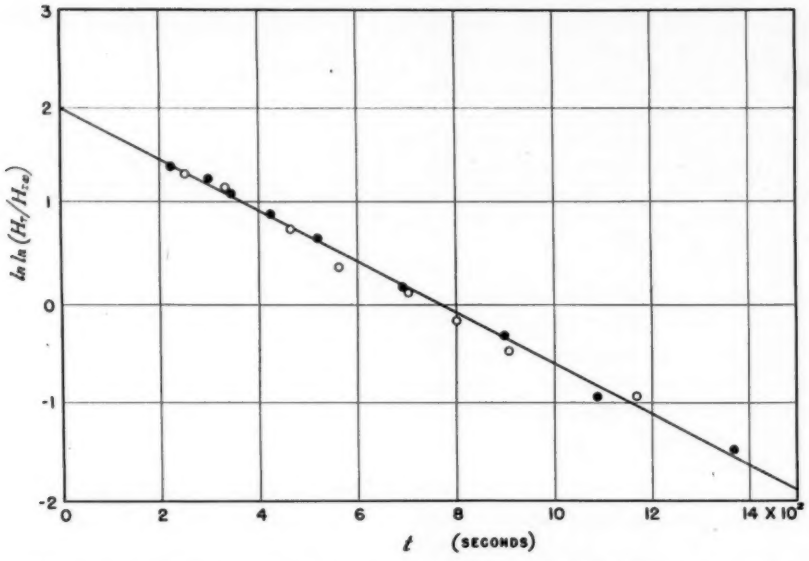


FIG. 3. Dark-adaptation. Rod vision, data of Wald and Clark (reference 27), observer A-B.C., natural pupil, left eye, light-adapted to $H_A=3330$ blondels. ○ 600-sec. exposure time, ● 1200-sec. exposure time.

It is a non-linear differential equation of the first order. Graphical or mechanical methods may be used in its solution even when the helios $H(s)$ is a function of time.

But usually in visual experiments, $H(s)$ is a constant over a light-adaptation period τ. In the remainder of this paper, we confine ourselves to the simple case where

(1) $H(s)=$const. during adaptation period τ,

(2) $\zeta(\delta, 2; s)=$const. during adaptation period,

(3) $\exp[-b(t-s)]\leqq 1$. This quantity is a weighting function whose purpose is to attenuate the effect of exposure to light that occurred in the remote past. Experiments indicate that unless values of an hour or more are considered, this weighting function is essentially equal to unity. When these three conditions are satisfied, Eq. (9) assumes the more amenable form,

$$\frac{dx}{dt}=k_1H(t)[\exp(-ax)-\exp(-a)]\zeta(\delta, 2; t)$$
$$-\{k_2-k_4x^3+A[1+ak_1H_A\zeta(\delta_A, 2)t]^{-e}\}x, \quad (9a)$$

where δ_A is the fixed pupil size and H_A the fixed helios during light adaptation.

Equilibrium

The general solution of Eq. (9a) is complicated, but simple solutions can be written in several important cases. An exact solution for the steady state is possible. Assume that conditions have remained constant for a sufficient time so that equilibrium obtains. Then $dx/dt=0$. Also $t\to\infty$, so $[1+ak_1H_A\zeta(\delta_A, 2)t]^{-e}\to 0$. Thus,

$$k_1H_A[\exp(-ax_A)-\exp(-a)]\zeta(\delta_A, 2)$$
$$=(k_2-k_4x_A^3)x_A,$$

and

$$H_A=\frac{(k_2-k_4x_A^3)ax_A}{ak_1[\exp(-ax_A)-\exp(-a)]\zeta(\delta_A, 2)}, \quad (10)$$

where x_A is the equilibrium concentration corresponding to the constant helios H_A. The equilibrium pupil diameter is δ_A.

Dark Adaptation

In dark-adaptation tests, the observer is exposed to a field of helios H_A for a time τ. At the end of this exposure time, the helios is suddenly reduced to zero and the eye begins to adapt itself to darkness, a process that takes approxi-

mately half an hour for the rods. The beginning of the dark-adaptation run is taken at the instant $t=0$, at which instant the concentration is $x=x_0$.

For the dark-adaptation run, $H(t)=0$. Thus Eq. (9a) reduces to

$$\frac{dx}{dt}=-\{k_2-k_4x^3 + A[1+ak_1H_A\zeta(\delta_A, 2)\tau]^{-c}\}x. \quad (11)$$

The solution of this differential equation is

$$t=\frac{1}{3B}\ln\left(\frac{x_0^3}{x^3}\frac{B-k_4x^3}{B-k_4x_0^3}\right), \quad (12)$$

where

$$B=k_2+A[1+ak_1H_A\zeta(\delta_A, 2)\tau]^{-c}. \quad (13)$$

In these equations, H_A is the helios to which the eye was exposed during the light-adaptation run of τ sec. During this period the pupil is of constant diameter δ_A. Equation (12) can be solved for ax:

$$ax=ax_0[\exp(3Bt) - k_4x_0^3(\exp(3Bt)-1)/B]^{-1}. \quad (12a)$$

The foregoing equation has expressed the concentration x as a function of time during the dark-adaptation run. The results can be expressed equally well in terms of threshold helios. According to Postulate IV,

$$f_t=k_3\left(\frac{dx}{dt}\right)_1=k_1k_3H_T(t)\zeta(\delta_T, 2; t) \times[\exp(-ax)-\exp(-a)],$$

or

$$H_T/H_{T\infty}=[\zeta(\delta_T, 2; t) \times[\exp(-ax)-\exp(-a)]]^{-1}, \quad (14)$$

where

$$H_{T\infty}=f_T/k_1k_3. \quad (15)$$

The quantity $H_{T\infty}$ is the threshold helios for a 2-mm pupil and $t\rightarrow\infty$. The values of ax from Eq. (12a) are substituted in Eq. (14) to give the threshold as a function of time.

Light Adaptation

The solution of Eq. (9a) for light adaptation will now be considered. In the usual laboratory test, the eye is dark adapted before the beginning of the light adaptation. At $t=0$, a large field of helios H_A is suddenly exposed and this helios remains constant for the duration of the run. The

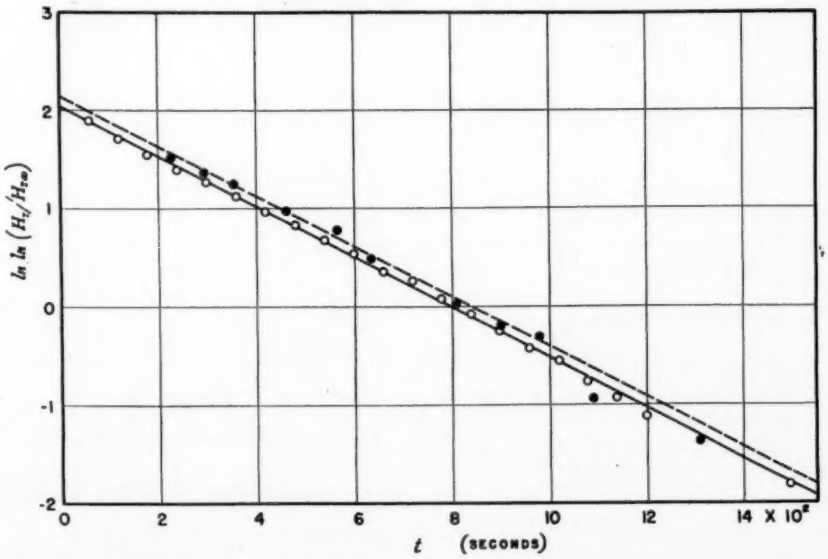


FIG. 4. Dark-adaptation curve for the honey bee. ○—○ Honey bee, data of Wolf and Zerrahn-Wolf (reference 28); ●--● Human observer, data of Hecht, Haig, and Chase (reference 29), observer S.H., $H=3800$, $\tau=120$ sec., rod vision.

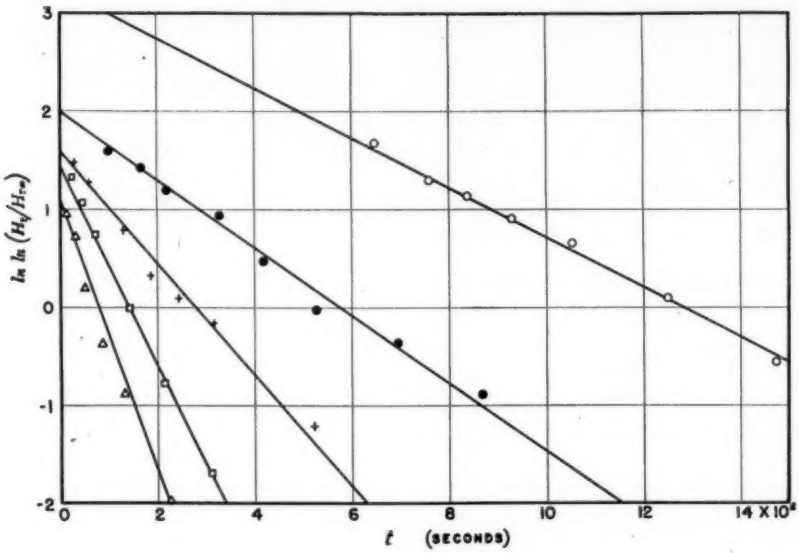


FIG. 5. Dark-adaptation curves showing the change of slope caused by altering the helios employed in light adaptation. Data of Haig (reference 30), rod vision, observer E.M.H., exposure time 240 sec., 2-mm fixed pupil. ○ $H_A = 47{,}000$ blondels; ● 4470; + 440; □ 200; △ 40.

initial condition is $x=0$ at $t=0$. Thus the solution of Eq. (9a) is

$$\int_0^x \frac{dx}{k_1 H_A[\exp(-ax)-\exp(-a)]\zeta(\delta_A, 2)-\{k_2-k_4x^3+A[1+ak_1H_A\zeta(\delta_A, 2)t]^{-c}\}x} = \int_0^t dt,$$

or

$$t=\int_0^x \{[k_1H_a[\exp(-ax)-\exp(-a)]\zeta(\delta_A, 2)-\{k_2-k_4x^3+A[1+ak_1H_A\zeta(\delta_A, 2)t]^{-c}\}x\}^{-1}dx. \quad (16)$$

The concentration x can be obtained as a function of t by performing this integration. The integration can be effected step by step, using graphical or numerical methods.

4. SPECIAL CASES

In many cases, the concentration x is small enough so that the equations are greatly simplified. These simplified equations apply always in cone vision and often in rod vision. If $x \ll 1$, the differential equation becomes

$$\frac{dx}{dt}=k_1H(t)\zeta(\delta, 2; t)\exp(-ax) - \{k_2+A[1+ak_1H_A\zeta(\delta_A, 2)t]^{-c}\}x. \quad (9b)$$

For *equilibrium*, the solution is obtained from Eq. (10) which becomes

$$H_A=\left[\frac{k_2}{ak_1\zeta(\delta_A, 2)}\right](ax_A)\exp(ax_A). \quad (10a)$$

For *dark adaptation*, Eq. (12a) reduces to the simple equation,

$$ax=ax_0\exp(-Bt). \quad (12b)$$

The threshold is obtained from Eq. (14), or

$$H_T/H_{T\infty}=\zeta^{-1}(\delta_T; 2; t)\exp(ax). \quad (14a)$$

Substitution of Eq. (12b) gives

$$\ln(H_T/H_{T\infty})=-\ln\zeta(\delta_T, 2; t) + ax_0\exp(-Bt). \quad (17)$$

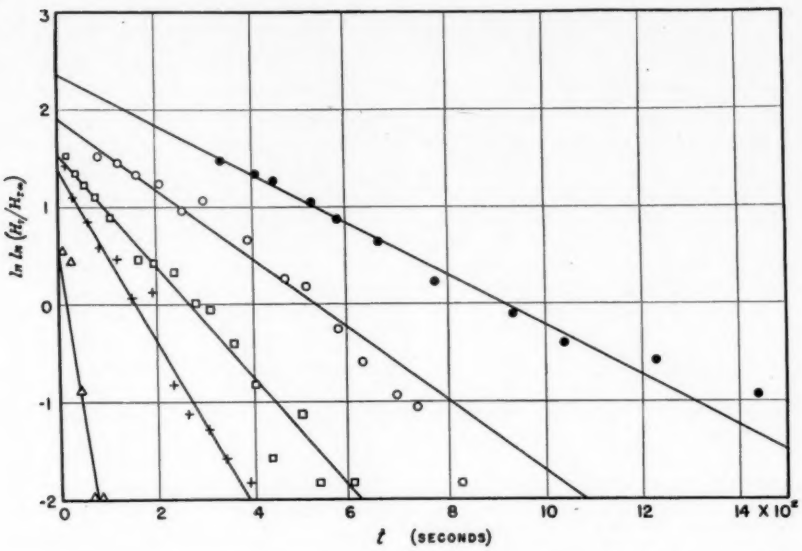


FIG. 6. Dark-adaptation curves showing change of slope caused by change in light adaptation. Data of Wald and Clark (reference 27), rod vision, observer A-B.C., natural pupil. ● Photoflash lamp; ○ $H=18{,}900$, $\tau=5$ sec.; □ $H=7500$, $\tau=5$; + $H=2450$, $\tau=5$; △ $H=36$, $\tau=5$.

For *light adaptation*, Eq. (16) becomes

$$t=\int_0^x \frac{\exp(ax)dx}{k_1H_A\zeta(\delta_A,2)}$$

$$=[ak_1H_A\zeta(\delta_A,2)]^{-1}[\exp(ax)-1].$$

Thus,

$$ax=\ln[1+ak_1H_A\zeta(\delta_A,2)t]. \qquad (16a)$$

The threshold is, from Eqs. (14a) and (16a),

$$H_T/H_{T\infty}=\zeta^{-1}(\delta_T,2;t)\cdot[1+ak_1H_A\zeta(\delta_A,2)t]. \quad (18)$$

Equation (16a) applies when x is very small (see Section 9). The range can be extended somewhat by expanding the integrand of Eq. (9b) in a series and integrating term by term. Then

$$t=[ak_1H_A\zeta(\delta_A,2)]^{-1}[\exp(ax)-1]$$

$$+[ak_1H_A\zeta(\delta_A,2)]^{-2}\left\{k_2\left(\frac{ax}{2}-\frac{1}{4}\right)\right.$$

$$\times\exp(2ax)+A\left(\frac{2ax}{3}-\frac{4}{9}\right)\exp(3ax/2)$$

$$\left.+\left(\frac{k_2}{4}+\frac{4A}{9}\right)\right\}+[ak_1H_A\zeta(\delta_A,2)]^{-3}$$

$$\times\left\{k_2^2\left[\frac{(ax)^2}{3}-\frac{2ax}{9}+\frac{2}{27}\right]\exp(3ax)\right.$$

$$+2k_2A\left[\frac{2(ax)^2}{5}-\frac{8ax}{25}+\frac{16}{125}\right]\exp(5ax/2)$$

$$+A^2\left[\frac{(ax)^2}{2}-\frac{ax}{2}+\frac{1}{4}\right]\exp(2ax)$$

$$\left.-\left[\frac{2k_2^2}{27}+\frac{32k_2A}{125}+\frac{A^2}{4}\right]\right\}+\cdots. \qquad (19)$$

Another special case of some importance occurs when $x\ll1$ and $\delta=2$. Then the differential equation is

$$\frac{dx}{dt}=k_1H(t)\exp(-ax)$$

$$-\{k_2+A[1+ak_1H_At]^{-e}\}x. \qquad (9c)$$

For *equilibrium*,

$$H_A=\left(\frac{k_2}{ak_1}\right)(ax_A)\exp(ax_A). \qquad (10b)$$

For *dark adaptation*,

$$B=k_2+A[1+ak_1H_A\tau]^{-e} \qquad (13a)$$

and

$$\ln(H_T/H_{T\infty})=ax_0\exp(-Bt). \qquad (17a)$$

BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

TABLE V. Analysis of dark-adaptation curves (rod vision).

Authority	H (blondel)	τ (sec.)	Measured $\log (H_{T\infty})$	Measured $\ln (ax_0)$	ax_0	B Measured	B Calculated
Haig, Table I[a]	4470	600	$\bar{4}.25$	2.45	11.6	4.35×10^{-3}	2.88×10^{-3}
	4470	360	$\bar{4}.25$	2.45	11.6	4.35	3.13
	4470	240	$\bar{4}.25$	2.45	11.6	4.35	3.39
	4470	120	$\bar{4}.25$	2.10	8.20	4.35	3.96
	4470	60	$\bar{4}.25$	1.80	6.05	5.50	4.77
	4470	24	$\bar{4}.25$	1.60	4.95	7.00	6.37
	4470	6	$\bar{4}.30$	1.30	3.67	9.60	10.7
Haig, Table III[a]	47,000	240	$\bar{4}.00$	3.25	25.8	2.52	2.14
	20,900	240	$\bar{5}.95$	2.75	15.6	2.54	2.64
	11,500	240	$\bar{5}.95$	2.35	10.5	2.70	2.87
	4470	240	$\bar{5}.90$	2.00	7.40	3.44	3.39
	440	240	$\bar{4}.00$	1.60	4.95	5.70	6.41
	200	240	$\bar{4}.00$	1.45	4.25	10.1	8.52
	40	240	$\bar{4}.00$	1.10	3.00	13.7	16.2

[a] See reference 30 of text.

For *light adaptation*,

$$ax=\ln[1+ak_1H_At], \tag{16b}$$

and

$$H_T/H_{T\infty}=1+ak_1H_At. \tag{18a}$$

A final case should be mentioned, where $x\ll1$, $\delta=2$, and light adaptation has been carried out thoroughly so that $H_A\tau\to\infty$. Then, from Eq. (13a), $B=k_2$ and

$$\ln(H_T/H_{T\infty})=ax_0\exp(-k_2t), \tag{17b}$$

or

$$\ln\ln(H_T/H_{T\infty})=\ln(ax_0)-k_2t. \tag{17c}$$

Such equations will be found useful later.

The *instantaneous threshold* is another quantity that has been studied experimentally. Procedure is the same as in dark-adaptation tests except that only one reading is taken for each value of H_A. The threshold H_{T0} is determined immediately after turning off the adapting light. The initial steepness that seems to be obtained with experimental dark-adaptation curves might lead one to doubt the existence of an instantaneous threshold. Calculation shows,[6,7] however, that the curves stay almost constant for an appreciable period after the adapting light is extinguished.

From Eq. (14), the instantaneous threshold is written

$$H_{T0}=\frac{(H_{T\infty})_2}{\zeta(\delta_T,2;0)[\exp(-ax_0)-\exp(-a)]}, \tag{14b}$$

where x_0 is the initial concentration of decomposition product and $(H_{T\infty})_2$ is the absolute threshold for a 2-mm pupil. For a 2-mm pupil and rod vision,

$$H_{T0}=\frac{5.5\times10^{-5}}{\exp(-ax_0)-\exp(-a)},$$

and for cone vision,

$$H_{T0}=24.5\times10^{-3}\exp(ax_0).$$

These equations may be used in calculating a curve of H_{T0} *vs* H, and the calculated curve will be found to bear a strong resemblance to the experimental one obtained by Blanchard.

5. AREA

To make the theory applicable to a larger group of conditions, two subsidiary phenomena must be considered—area of stimulus, and pupil size. The former will be treated in this section.

In previous work,[7] a simplification was effected by assuming that the test object was so large that variations in its area had negligible effect. Such a simplification, however, rules out a number of interesting phenomena which can now be considered. It is well known that the lowest helios threshold is obtainable when the test object subtends a large solid angle ω at the eye and thus stimulates a large number of receptor elements (rods or cones). As ω is decreased, the threshold rises. We wish to obtain the equation relating threshold helios H_T and solid angle ω subtended at the eye by the test spot.

The problem is not a simple one. It depends on the rod and cone densities and the variation

of these densities over the retina.[10] It is impossible to formulate the problem rigorously in terms of ω, since the number of excited elements depends on the shape, the orientation, and the position of the image on the retina. To obtain a solution that is simple enough to be practicable, however, we shall make the simplifying assumption that ω is the only independent variable and that all cases can be classified as

(*a*) Cone vision, central fixation, circular test spot; or

(*b*) Rod vision, fixation at some distance from the fovea, uniform rod density assumed over the area considered.

With the simplified conditions, one recognizes immediately two limiting cases for either (*a*) or (*b*):

(1) Very large ω. The number of stimulated receptors is so large that a reasonable variation in their number causes little change, and

$$H_T \leqq \text{const.} \tag{20}$$

(2) Very small ω. If the test object is so small that it cannot be resolved by the eye, helios alone has no significance and only flux density is effective:

$$H_T\omega \leqq \text{const.} \tag{21}$$

Between these two extreme cases must lie a transition region where there is a gradual change from one of the foregoing equations to the other.

Many researches have been devoted to the study of area. These investigations are characterized by the small range of ω and the stupid ignoration of the foregoing classes (1) and (2). One of the best-known results is that of Riccò[11] (1877), who found that Eq. (21) held for foveal vision, a result that is often known as *Riccò's law*. The validity of this law has been hotly contested by many who have used extra-foveal vision and a narrow range of ω differing from that used by Riccò. Thus Piper[12] (1903) claimed that

$$H_T = \frac{c}{\omega^{\frac{1}{2}}}, \tag{22}$$

where c is a constant and ω appears as the $\frac{1}{2}$ power instead of the first power. Henius[13] (1909) and Fujita[14] showed that Eq. (22) is valid only over a narrow range of ω. Abney and Watson[15] (1916) used the equation,

$$H_T = \frac{c}{\omega^k}, \tag{23}$$

where $k=0.85$, while Piéron[16] (1920) employed the same equation with $k=0.88$. More recently, Wald[17] has advocated the expression,

$$H_T = \frac{c}{(\omega-\omega_0)^k}, \tag{24}$$

where c, ω_0, and k are experimentally determined constants.

All of these equations are satisfactory and any

TABLE VI. Slopes of dark-adaptation curves. (Data plotted in Fig. 7).

Authority	H (blondel)	τ (sec.)	$(1+khH\tau)$	B (measured)	$(B-2.00\times10^{-3})$
Haig, Table I[a]	4470	600	4021	4.35×10^{-3}	2.35×10^{-3}
	4470	360	2421	4.35	2.35
	4470	240	1611	4.35	2.35
	4470	120	806	4.35	2.35
	4470	60	404	5.50	3.50
	4470	24	162	7.00	5.00
	4470	6	41.3	9.60	7.60
Haig, Table III[a]	47,000	240	16,940	2.52	0.52
	20,900	240	7530	2.54	0.54
	11,500	240	4140	2.70	0.70
	4470	240	1611	3.44	1.44
	440	240	159	5.70	3.70
	200	240	73	10.1	8.1
	40	240	15.4	13.7	11.7

[a] See reference 30 of text.

TABLE VII. Numerical constants.

Quantity	Rods	Cones
ak_1	1.50×10^{-3}	3.0×10^{-5}
k_2	2.00×10^{-3}	0.0110
A	55.6×10^{-3}	0.945
c	0.500	1.20
$H_{T\infty}$	3.6×10^{-6}	3.0×10^{-3}
$(H_{T\infty})_2$	55×10^{-6}	24.5×10^{-3}
af_T/k_3	8.25×10^{-8}	7.35×10^{-7}
a	9.0	30
k_1	1.67×10^{-4}	1.0×10^{-6}
k_4	1.98×10^{-3}	0
A_3	0.0469	0.00874

[10] G. Østerberg, Acta Ophthal. Suppl. **6**, 1 (1935).

[11] Riccò, Ann. ottal. **3**, 373 (1877).

[12] H. Piper, Psychol. u. Physiol. Sinnesorgane **31**, 161 (1903).

[13] K. Henius, Zeits. f. Sinnesphysiol. **43**, 99 (1909).

[14] T. Fujita, Zeits. f. Sinnesphysiol. **43**, 243 (1909).

[15] W. Abney and W. Watson, Phil. Trans. Roy. Soc. London **216**, 91 (1916).

[16] H. Piéron, Comptes rendus Soc. Biol. **83**, 753 (1920).

[17] G. Wald, J. Gen. Physiol. **21**, 269 (1938).

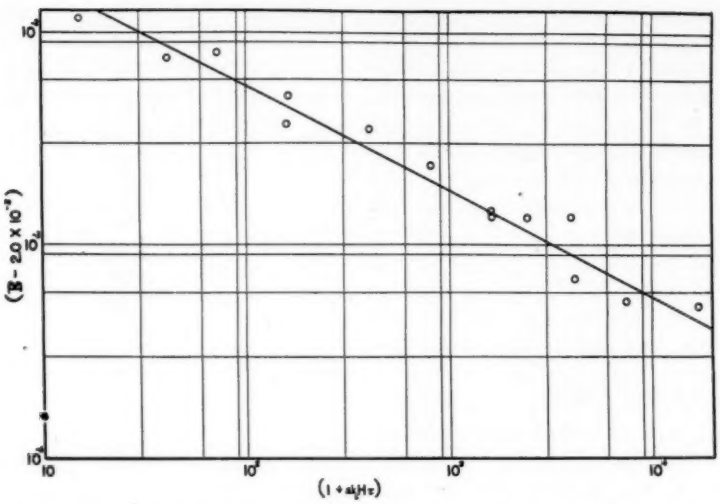


FIG. 7. Slopes of dark-adaptation curves. The data are from Table VI and the curve determines the constants A and c in Eq. (13a).

value of k from 0 to 1 may be employed if one takes a narrow enough range of ω, selected at the proper absolute value. But notice that none of these proposed equations fits the two required special cases, Eqs. (20) and (21). A satisfactory equation is not hard to find. A number of equations satisfy all requirements, including reduction to the special cases (1) and (2). Perhaps the simplest of these allowable equations is of the form (Postulate V),

$$H_T/H_{T\infty}=\frac{1}{\omega}-[A_3+\omega^{\frac{1}{2}}]^2. \qquad (25)$$

For very small angles, $\omega^{\frac{1}{2}}$ becomes negligible and the threshold varies inversely as ω, in accordance with Riccò's law. For very large angles, A_3 becomes negligible and H_T reaches the absolute threshold $H_{T\infty}=\text{const.}$

For cone vision and central fixation, all reliable data show that Riccò's law is approximately true if the image of the test object falls within the fovea. But a circular disk covering the fovea subtends a solid angle of only about 5×10^{-4} steradian. A larger disk will encroach on the parafovea, where the cone density is less[10] and where a departure from Riccò's law might be expected. The form of curve, based on some data of Steinhardt,[18] is shown in Fig. 1. The experimental data are in good agreement with the curve, which is plotted from Eq. (25) with $A_3=0.00874$. Since logarithmic scales are employed along both axes, the curve extends indefinitely upward to the left with a slope of -1.

This straight portion of the curve is shown to better advantage in Fig. 2, which applies to rod vision. Data for a wide range of ω were obtained by Reeves.[19] Data of Houston[20] and of Wald[17] are also included.[21] The dotted line represents Riccò's law, which begins to fail at approximately $\omega=10^{-4}$ steradian. The continuous curve is plotted from Eq. (25) with $A_3=0.0469$. The good agreement of the equation with Wald's data is indicated by Table I.

We conclude, therefore, that the size of test object affects the threshold, over the complete range of ω, according to the equation,

$$H_T/H_{T\infty}=\frac{1}{\omega}-[0.00874+\omega^{\frac{1}{2}}]^2 \qquad (26)$$

for *cone vision*, and

$$H_T/H_{T\infty}=\frac{1}{\omega}-[0.0469+\omega^{\frac{1}{2}}]^2 \qquad (27)$$

for *rod vision*.

[18] J. Steinhardt, J. Gen. Physiol. **20**, 185 (1936).

[19] P. Reeves, Astrophys. J. **47**, 141 (1918).

[20] R. A. Houston, *Vision and Color Vision* (Longmans, Green and Company, London, 1932), p. 50.

[21] W. J. Crozier and A. H. Holway, J. Gen. Physiol. **23**, 101 (1939).

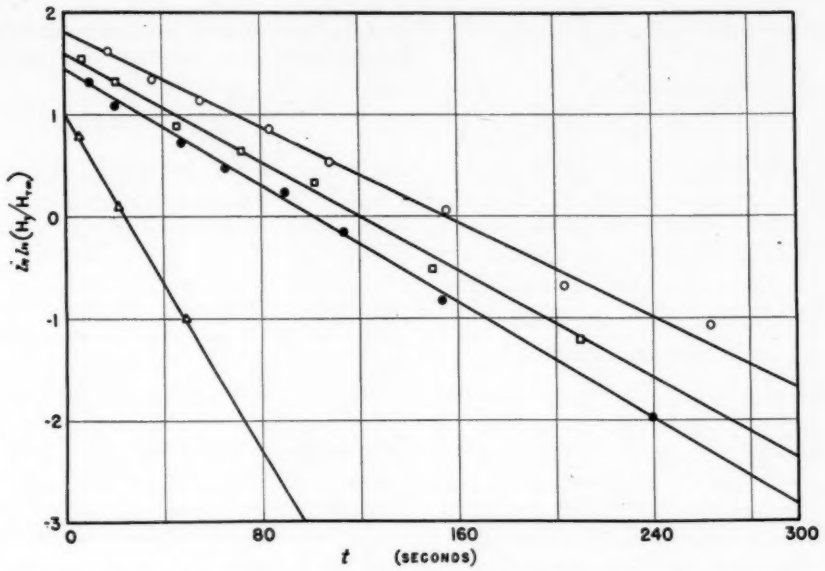


FIG. 8. Dark-adaptation curves for cones. Data of Hecht, Haig, and Chase (reference 29). Observer A.M.C., red light, exposure time 120 sec. ○ $H=400{,}000$ blondels; □ 38,900; ● 19,500; △ 3800.

6. PUPIL DIAMETER

Another factor that must be considered is pupil diameter. The velocity of the photochemical reaction in a given receptor element depends on the luminous flux to that element. But no direct measurements of this quantity can be made, and one must content himself with measurements of the helios of the surface which is imaged on the retina. If a fixed pupil is employed, a definite ratio exists between the retinal flux and the external helios; but if the natural pupil is used, its diameter must be known under all circumstances.

For equilibrium conditions, the average pupil diameter is expressed by the equation,[6]

$$\delta=4.90-3.00\tanh\left[0.400\,(\log H-0.50)\right] \quad (28)$$

where H=external helios (blondel), δ=pupil diameter (mm).

The *Stiles factor*[6] relates two values of helios—a value H_δ obtained with natural pupil of diameter δ, and a value H_β which will give the same visual effect when used with an artificial pupil of diameter β. The relation is

$$\zeta(\delta,\beta)=H_\beta/H_\delta. \quad (29)$$

For rod vision, this helios ratio is inversely proportional to the ratio of the pupil areas, or

$$\zeta(\delta,\beta)=(\delta/\beta)^2. \quad (30)$$

For cone vision, the relation is more complicated because of the presence of the Stiles-Crawford effect,[22] and

$$\zeta(\delta,\beta)=\left(\frac{\delta}{\beta}\right)^2\frac{\left[1-A\delta^2/8+B\delta^4/48\right]}{\left[1-A\beta^2/8+B\beta^4/48\right]}. \quad (31)$$

For the case of $\beta=2$,

$$\zeta(\delta,2)=0.2620\delta^2\left[1-1.06\times10^{-2}\delta^2+4.17\times10^{-5}\delta^4\right]. \quad (32)$$

For equilibrium conditions, the values of δ to be used in Eqs. (30) and (31) are obtained from Eq. (29). For transient conditions, experiments show that[6]

$$\frac{\delta-\delta_{\min}}{\delta_{\max}-\delta_{\min}}=1.10\exp(-t), \quad (33)$$

[22] W. S. Stiles and B. H. Crawford, Proc. Roy. Soc. **B112**, 428 (1933).

for *closing of the pupil*, and

$$\frac{\delta_{max}-\delta}{\delta_{max}-\delta_{min}}=0.880\exp(-0.228t)+0.160\exp(-0.0125t) \quad (34)$$

for *opening of the pupil*. These equations apply to any range $(\delta_{max}-\delta_{min})$ of pupil size, these two limits being calculated from the known values of helios, in accordance with Eq. (30). The effect of pupil size is discussed in greater detail in two previous papers.[6]

7. THE ABSOLUTE THRESHOLDS

It has long been known that a visual threshold exists, below which there is no sensation. In Section 5, two limiting cases of this threshold are considered; one obtained with very large test objects and giving $H_T=$const., and the other with very small test objects and with $H_T\omega=$const. In either case, the threshold is obtained with dark-adapted eyes. After 30 minutes or more in the dark, the observer determines the weakest light that can be detected. Consider first the case for very small sources. For 2000 years a star classification into 6 "magnitudes" has been used. A 6th-magnitude star is the weakest that can be detected by the unaided eye. According to Russell,[23] the star magnitude m is related to flux density by the equation,

$$m=-14.18-2.512\log D, \quad (35)$$

where $D=$luminous flux density (lumen m^{-2}) on a surface at the eye and normal to the line of sight. Thus a 6th-magnitude star produces a flux density D of 85×10^{-10} lumen m^{-2} at the eye. According to Ives,[24] this value corresponds to a radiant flux density of 1.35×10^{-11} watt m^{-2} or to 38.0×10^{-17} watt through a pupil of 28 mm^2 area.

But the foregoing values do not represent the real threshold. The background of the night sky is by no means of zero helios, and lower thresholds are found with artificial stars viewed in the laboratory with a completely dark background. Under these conditions, a flux density of approximately 10×10^{-10} lumen m^{-2} can be detected, corresponding to approximately 8×10^{-17} watt through a pupil of 8-mm diameter.

A summary of experimental results is given in Table II. The Ives value of $D_T=85\times10^{-10}$

[23] H. N. Russell, Astrophys. J. **41**, 103 (1916).
[24] H. E. Ives, Astrophys. J. **44**, 124 (1916).

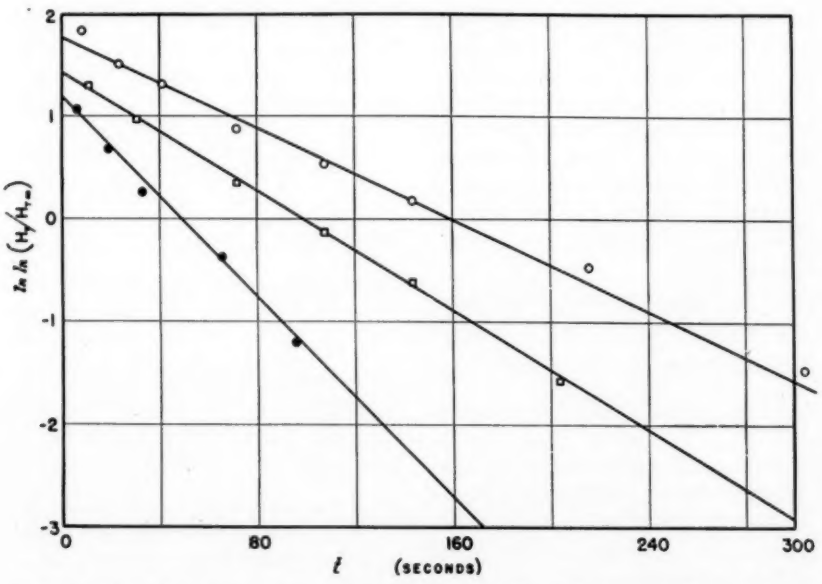


FIG. 9. Dark-adaptation curves. Cone vision, data of Hecht, Haig, and Chase (reference 29), observer A.M.C., violet light, exposure time 120 sec. The symbols refer to the same values of helios used in Fig. 8.

TABLE VIII. Analysis of dark-adaptation curves (cone vision). Data of Hecht, Haig, and Chase.[a] $\tau = 120$ sec.

Observer	Light	H (blondel)	log ($H_{T\infty}$)	Measured ln (ax_0)	ax_0	B Measured	B Calculated
S.H.	Violet	400,000	$\bar{1}.73$	1.85	6.36	1.11×10^{-2}	1.15×10^{-2}
		38,900	$\bar{1}.73$	1.17	3.22	1.46	3.09
		19,500	$\bar{1}.70$	1.18	3.25	2.00	3.02
	Red	400,000	$\bar{1}.82$	1.97	7.20	1.30	1.12
		38,900	$\bar{1}.97$	1.71	5.54	1.42	1.22
		19,500	$\bar{1}.92$	1.52	4.57	1.62	1.49
		3,800	$\bar{0}.01$	1.00	2.72	5.20	4.70
C.H.	Violet	400,000	$\bar{1}.54$	1.50	4.48	1.33	1.53
		38,900	$\bar{1}.57$	1.40	4.06	1.72	1.83
		19,500	$\bar{1}.54$	1.12	3.07	2.36	3.48
	Red	400,000	0.02	1.64	5.15	0.82	1.30
		38,900	0.00	1.50	4.48	2.14	1.53
		19,500	0.09	1.25	3.49	4.15	2.53
		3,800	0.10	1.25	3.49	4.15	2.53
A.M.C.	Violet	400,000	$\bar{1}.51$	1.78	5.92	1.13	1.19
		38,900	$\bar{1}.61$	1.42	4.14	1.45	1.76
		19,500	$\bar{1}.70$	1.20	3.32	2.44	2.85
	Red	400,000	$\bar{1}.66$	1.81	6.10	1.16	1.16
		38,900	$\bar{1}.65$	1.60	4.95	1.32	1.35
		19,500	$\bar{1}.67$	1.45	4.26	1.43	1.67
		3,800	$\bar{1}.62$	1.00	2.72	4.11	4.70

[a] See reference 29 of text.

lumen m^{-2} is based on a star of 6th magnitude. It corresponds to a candle at a distance of 10.8 km in a perfect atmosphere. When the disturbing effect of sky helios is eliminated, a star magnitude between 7.4 and 8.6 is possible, corresponding to a candle at 20 to 34 km.

The absolute helios threshold is determined by exposing a luminous disk subtending a large visual angle (diameter subtending an angle of 7° or greater) in a dark room. Table III gives some of the available data for rod vision. Table IV is for cone vision. The values for rods cover a range of about 10:1, which is not surprising in view of the 100:1 variation obtainable on a given individual by a deficiency in vitamin A.[25] Representative values of the absolute helios threshold will be taken as

$$H_T = 3.6\times10^{-6} \text{ blondel for rods,}$$
$$H_T = 3\times10^{-3} \text{ blondel for cones.}$$

The assumption is usually made that there is no connection between the star threshold of Table II and the helios threshold of Table III. But in Section 5 we showed that these two thresholds are merely special cases of a general relationship between H_T and ω. Thus by fixing either threshold it is now possible to calculate the other by means of Eqs. (26) and (27).

By the definition of helios,[26]

$$H = \pi D/\omega, \tag{36}$$

and thus the flux-density threshold (star threshold) is

$$D_T = H_T\omega/\pi. \tag{36a}$$

For rod vision,

$$H_T/H_{T\infty} = \frac{1}{\omega}[0.0469+\omega^{\frac{1}{2}}]^2,$$
$$H_{T\infty} = 3.6\times10^{-6} \text{ blondel.} \tag{27}$$

When $\omega\to0$, as for a star, Eq. (27) reduces to

$$H_T = \frac{3.6\times10^{-6}}{\omega}(0.0469)^2,$$

and substitution in Eq. (36a) gives the flux density threshold,

$$D_T = \frac{3.6\times10^{-6}}{\pi}(0.0469)^2 = \mathbf{25.2\times10^{-10}\ lumen\ m^{-2}}.$$

Comparison with Table II shows that this calculation is in agreement with the value obtained by Reeves for the average of 3 observers.

[25] S. Hecht and J. Mandelbaum, J. Am. Med. Assn. **112**, 1910 (1939); Am. J. Physiol. **130**, 651 (1940). C. Sheard, J. Opt. Soc. Am. **34**, 464 (1944).

[26] P. Moon, J. Opt. Soc. Am. **32**, 348 (1942); Am. Phys. **10**, 134 (1942).

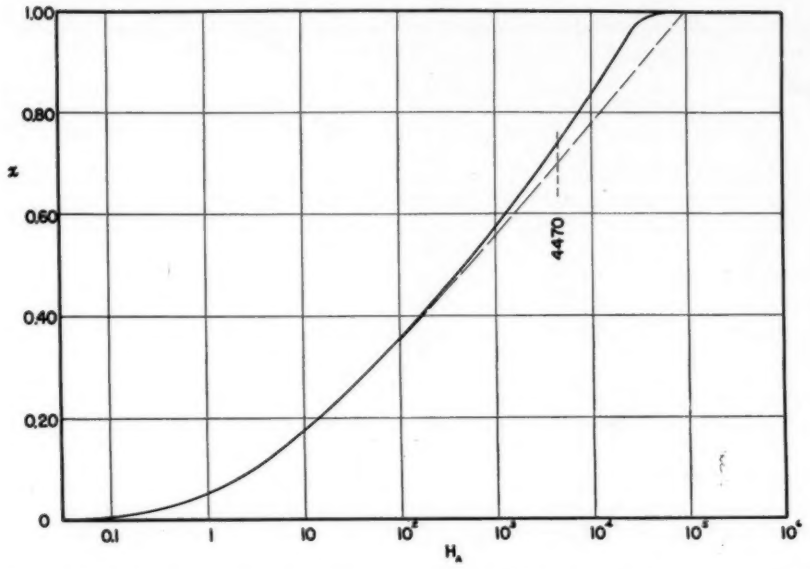


FIG. 10. Calculated concentration of retinene in a rod. Equilibrium condition, 2-mm pupil, $a=9$. — Eq. (10); -- Eq. (10b).

For cone vision,

$$H_T/H_{T\infty}=\frac{1}{\omega}[0.00874+\omega^{\frac{1}{2}}]^2, \qquad (26)$$

$$H_{T\infty}=3\times10^{-3} \text{ blondel.}$$

If $\omega\to0$, Eq. (26) becomes

$$H_T=\frac{3\times10^{-3}}{\omega}(0.00874)^2$$

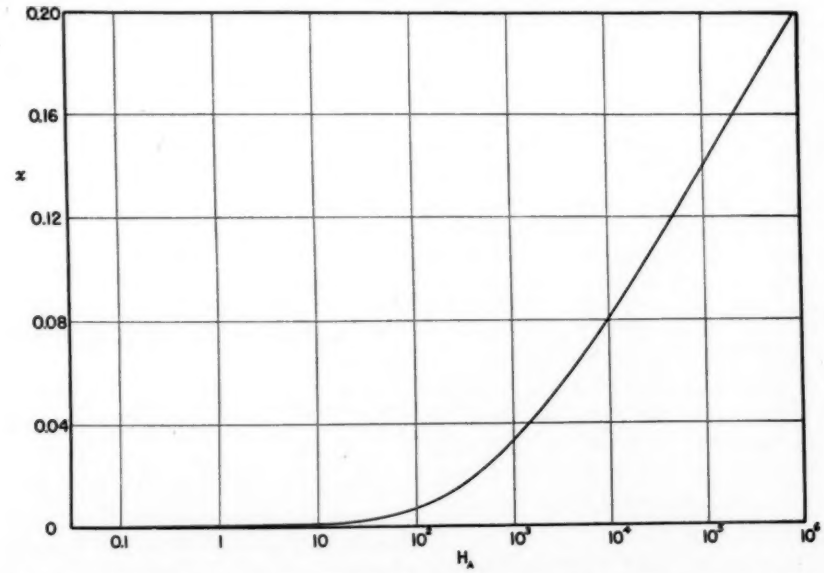


FIG. 11. Calculated concentration. Cone vision, equilibrium, 2-mm pupil, $a=30$, Eq. (10b).

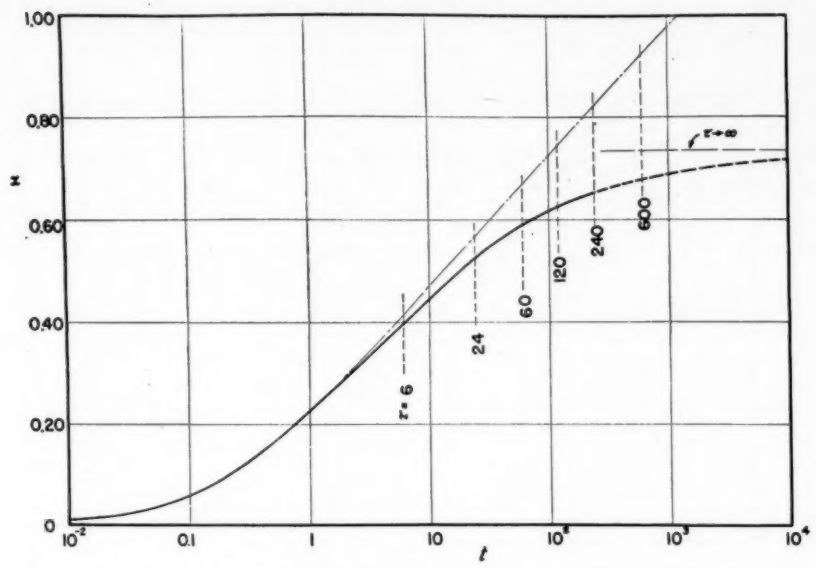


FIG. 12. Calculated light-adaptation curve giving the concentration of retinene in a rod as a function of time. $H=4470$. — Eq. (16); —·— Eq. (16b).

and

$$D_T=\frac{3\times10^{-3}}{\pi}(0.00874)^2=730\times10^{-10}\ \text{lumen m}^{-2}.$$

This calculation shows that the rods are approximately 30 times as sensitive as the cones in the detection of stars. We have not found any experimental data on star thresholds for cones, and therefore have no check on this calculated value.

Note that we are no longer confined to two special thresholds occurring at the two extremes of size. The equations allow the calculation of the threshold for ω from 0 to 2π. Moreover, the threshold can be expressed either as a helios of the test object or as an equivalent flux at the eye.

For rod vision,

$$H_T=\frac{3.6\times10^{-6}}{\omega}[0.0469+\omega^{\frac{1}{2}}]^2,$$
$$D_T=1.15\times10^{-6}[0.0469+\omega^{\frac{1}{2}}]^2. \tag{37}$$

For cone vision,

$$H_T=\frac{3.0\times10^{-3}}{\omega}[0.00874+\omega^{\frac{1}{2}}]^2,$$
$$D_T=0.95\times10^{-3}[0.00874+\omega^{\frac{1}{2}}]^2. \tag{38}$$

8. CONSTANTS

In Sections 3 and 4, equations were developed for the photochemical theory of vision. The equations apply to both rod vision and cone vision but the constants are different in the two cases. These constants must now be evaluated.

(*a*) In Section 4, it was shown that except at high values of x and small values of time the dark-adaptation curves for a 2-mm pupil can be represented by Eq. (17a). Taking the natural logarithms of both sides, one obtains

$$\ln\ln(H_T/H_{T\infty})=\ln ax_0-Bt. \tag{39}$$

According to this equation, a plot of *ln ln* $(H_T/H_{T\infty})$ *against* t *should be a straight line.* The line should have a negative slope of B and should intersect the axis ($t=0$) at a value $\ln ax_0$.

TABLE IX. Calculated values of time (sec.). $H=4470$ blondels.

x	Eq. (19) One term	Two terms	Three terms	Numerical integration
0.30	2.070	2.17	2.18	2.19
0.40	5.309	5.98	6.07	6.08
0.50	13.28	16.7	17.8	18.2
0.60	32.87	51.1	63.8	73.5
0.65	51.63	244	287	225

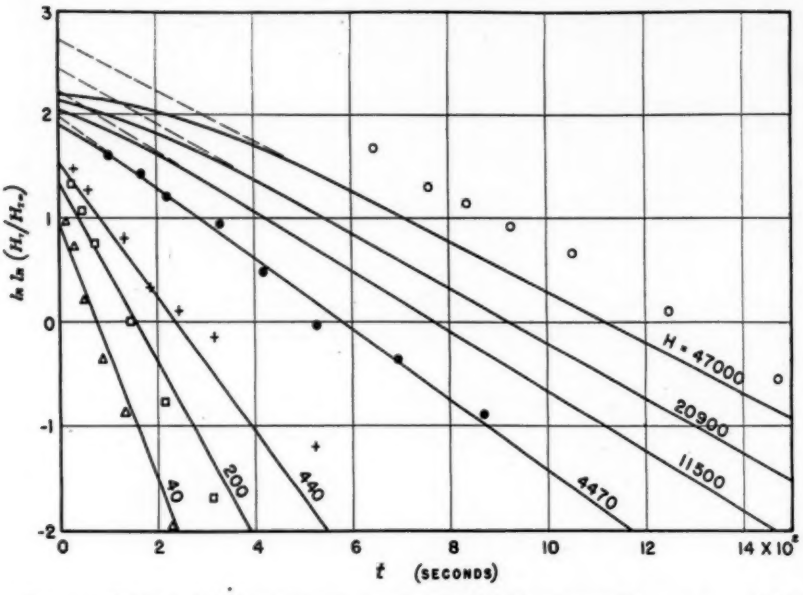


FIG. 13. Calculated dark-adaptation curves. The initial concentrations x_0 are obtained from Fig. 12. The experimental points are those of Haig (reference 30). Rod vision, observer E.M.H., exposure time 240 sec., 2-mm fixed pupil, symbols same as in Fig. 5.

That the predicted shape of curve agrees with experiment is shown by Fig. 3. The points represent experimental data of Wald and Clark.[27] No definite systematic deviation from the straight line can be detected, showing that the postulates of Section 2 lead to results in agreement with experiment. Figure 4 exhibits two other curves, one for the honey bee[28] and the other for a human observer.[29] Both sets of data are in good agreement with Eq. (39), as are many other curves shown in another paper.[7]

(*b*) From Eq. (16b) for an initial concentration $x=x_0$ after an exposure of $t=\tau$ sec.,

$$(ak_1)=\frac{1}{H\tau}[\exp(ax_0)-1]. \qquad (40)$$

Equation (40) can be employed in calculating the constant (ak_1). A representative dark-adaptation curve from Haig's research[30] gives

$$H\tau=1.072\times10^6, \quad ax_0=7.40.$$

Substitution in Eq. (40) evaluates ak_1 for rod vision:

$$ak_1=1.50\times10^{-3}.$$

(*c*) The foregoing value of (ak_1) is now substituted in Eq. (13a) which represents the slope of the dark-adaptation curves when plotted as $\ln\ln(H_T/H_{T\infty})$ *vs.* t. According to Eq. (13a), the slope B should be a function of the previous exposure time τ and the light-adaptation helios H_A. That this prediction of the theory is in accord with experiment is shown by the data of Figs. 5 and 6. The experimental results are again in agreement with straight lines, and the slopes of the lines become progressively greater as the light adaptation is reduced.

According to Eq. (13a), a plot of $\log(B-k_2)$ against $\log(1+ak_1H\tau)$ should be a straight line. All the Haig data (Table V) were taken to check this prediction. A few trials showed that if $k_2=2.00\times10^{-3}$, the data fitted a straight line as predicted by Eq. (13a). The graph shown in Fig. 7 is plotted from the data of Table VI. From this graph, the remaining constants are found to be

$$A=55.6\times10^{-3}, \quad c=0.500.$$

[27] G. Wald and A-B. Clark, J. Gen. Physiol. **21**, 93 (1937).
[28] E. Wolf and G. Zerrahn-Wolf, J. Gen. Physiol. **19**, 229 (1935).
[29] Hecht, Haig, and Chase, J. Gen. Physiol. **20**, 831 (1937).
[30] C. Haig, J. Gen. Physiol. **24**, 735 (1941).

Thus for rod vision, the slope is

$$B=[2.00+55.6(1+1.5H_T\times10^{-3})^{-1}]10^{-3}. \quad (13b)$$

(*d*) From Eq. (14), the frequency of nerve impulses is

$$f_T=k_1k_3(H_{T\infty})_2,$$

or

$$(af_T/k_3)=(ak_1)(H_{T\infty})_2 \quad (15a)$$

where $(H_{T\infty})_2$ is the absolute threshold ($t\rightarrow\infty$, $\omega\rightarrow2\pi$) for 2-mm pupil. For rod vision, the new constant (af_T/k_3) is

$$(af_T/k_3)=1.50\times10^{-3}(3.6\times10^{-6})15.3$$
$$=8.25\times10^{-8}.$$

(*e*) Most of the work can be done without a knowledge of a or of k_4. For rod vision it was found by trial and error that satisfactory values were

$$a=9.0, \quad k_4=1.98\times10^{-3}.$$

The various constants are listed in Table VII.

The foregoing evaluation has been for rod vision. For cone vision, the data of Hecht, Haig, and Chase were employed. Figures 8 and 9 show representative data.

(*a*) The straight-line representation, Eq. (39), is again applicable and the slopes increase as the adaptation helios H_A is decreased. A summary of the data is made in Table VIII.

(*b*) From Eq. (40) and a datum obtained with red light by A.M.C., the constant (ak_1) is evaluated. In the test, $H_A=19{,}500$, $\tau=120$, $ax_0=4.26$, so

$$(ak_1)=3\times10^{-5}.$$

(*c*) The slopes B are now considered as a function of $(1+ak_1H_T)$ and it is found that for cones,

$$k_2=0.0110, \quad A=0.945, \quad c=1.20.$$

(*d*) The constant (af_T/k_3) is determined from Eq. (15a):

$$(af_T/k_3)=3.0\times10^{-5}(3.0\times10^{-3})8.17$$
$$=7.35\times10^{-7}.$$

(*e*) For cone vision, the experimental data show no signs of saturation and the approximate equations of Section 4 apply under all circumstances. In these equations x always occurs in the product ax, so there is no necessity of evaluating a separately. However, a value can be

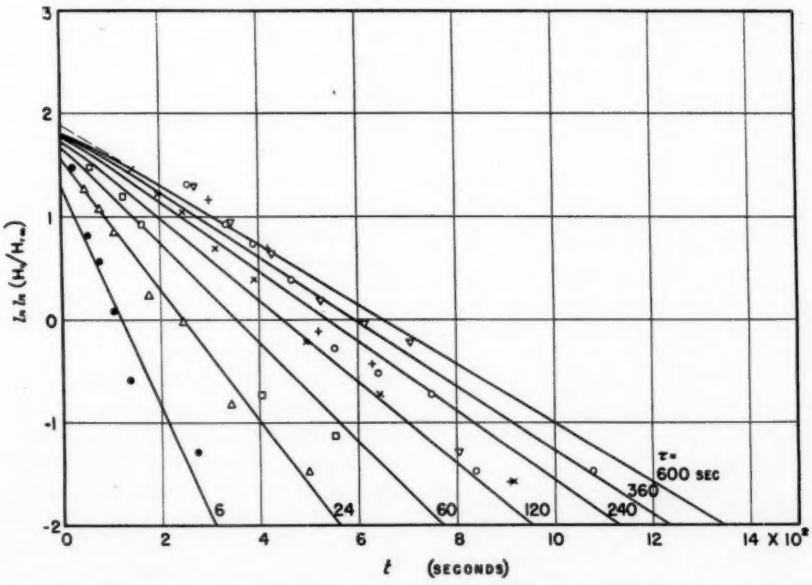


FIG. 14. Calculated dark-adaptation curves. Data of Haig (reference 30). Rod vision, observer E.M.H., exposure time 240 sec., 2-mm fixed pupil, $H_A=4470$. + 600 sec.; ▽ 360; ○ 240; × 120; □ 60; △ 24; ● 6.

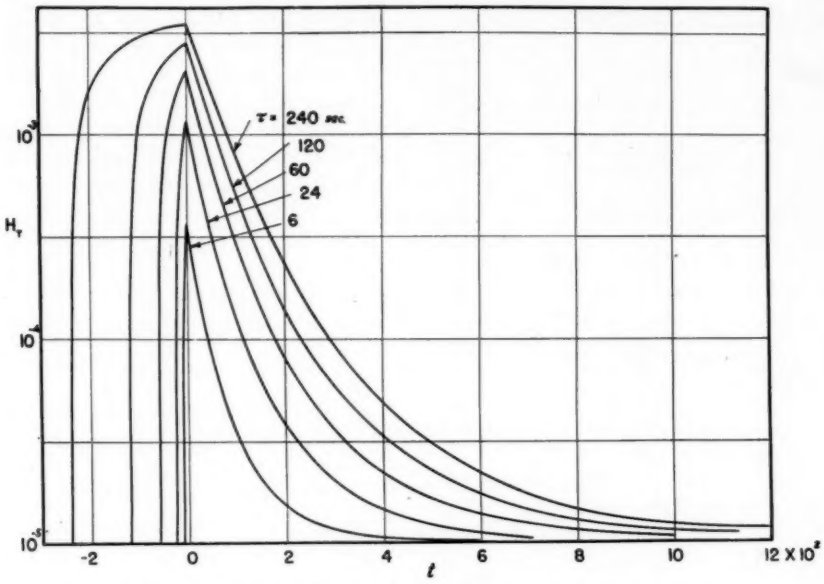


FIG. 15. Calculated light-adaptation and dark-adaptation curves. $H_A=4470$, 2-mm fixed pupil, rod vision.

given to a if desired. We have chosen $a=30$, which is sufficiently high so that saturation effects do not enter.

9. CONCENTRATIONS

The equations of the photochemical theory were developed in Sections 3 and 4, and the constants were evaluated in Section 8. Thus we are in a position to compute numerous visual phenomena. Figure 10 shows the computed curve of concentration x as a function of the adapting helios H_A. Results are for equilibrium with rod vision, and the calculations were made by means of Eq. (10). The dotted curve is an approximation obtained from Eq. (10b). This approximation is seen to be satisfactory for x less than about 0.4.

A similar curve for cone vision is shown in Fig. 11. Note that this approximate formula is satisfactory to beyond sunlight values. The difference in the appearance of the two curves is caused largely by the difference in a for rods and cones. With $a=9$ in Fig. 10, saturation is reached at $H \leqq 10^5$ blondels. But with $a=30$ in Fig. 11, no trace of saturation is noticeable, which is in accordance with experiments that indicate no saturation at $H=4\times10^5$.

Transient phenomena can also be computed. Figure 12 indicates the way that the concentration builds up under the action of a constant helios. The eye is initially dark-adapted, and at $t=0$ a helios of 4470 blondels is suddenly applied. The heavy curve is computed by numerical methods from Eq. (16). The computations were carried out to about 240 sec., beyond which the curve is estimated (dotted curve). The upper curve is an approximation calculated by means of Eq. (16a). The approximation is very easy to use but it is inexact for $x>0.3$. A comparison of the various equations is given in Table IX. The column on the right indicates the values of time calculated by numerical integration from Eq. (16). The second column gives the results obtained with Eq. (16a). The remaining columns refer to the series expansion, Eq. (19). The first term of this expansion is identical with Eq. (16a). The values of Table IX refer only to the specific case of $H=4470$. In general it is found[7] that Eq. (16a) is applicable if $\tau \leqq 5$ sec.

10. THRESHOLDS

A knowledge of the concentration allows one to calculate the threshold helios H_T under a

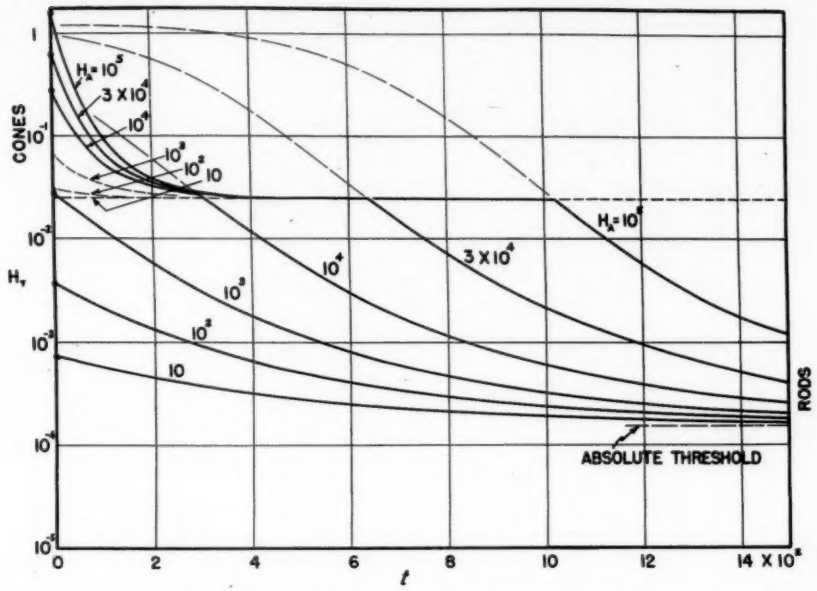


FIG. 16. Calculated dark-adaptation curves for both rods and cones. 2-mm pupil, $\tau \to \infty$. —— Final curve for eye; — — Rod vision, not employed because cones are more sensitive; ---- Cone vision, not employed because rods are more sensitive.

variety of conditions. The absolute threshold as function of size has already been considered. As a special case of the general equation of Section 5, we have Riccò's law. If the observer is dark-adapted but the threshold is determined for a single flash of light rather than for a continuous exposure, the equations give Bloch's law:

$$H_T \Delta t = \text{constant}. \qquad (41)$$

The exposure time $\Delta t \leq 0.1$ sec.

Dark-adaptation curves are among the most interesting applications of the photochemical theory. As noted in connection with Figs. 3 and 4, the form of curve predicted by the theory is in good agreement with experiment. The variation in slope predicted by the theory is also in agreement with the latest experimental data. A comparison of measured and calculated slopes is given in Table V. When the concentration x becomes rather high, the x^3 term in Eq. (9) becomes important and destroys the linearity of the dark-adaptation curves. Figure 13 gives a family of computed curves for constant exposure time of 240 sec. and for various values of adaptation helios. The effect of the x^3 term is noticeable only in the curves for $H_A \geqq 4470$ and only in these curves at small values of time. The initial concentrations and thresholds were calculated from the known values of helios. The slopes were also calculated, using Eq. (13). The only bad discrepancy between calculation and experiment is in the curve for $H = 47{,}000$. Even this could be made to fit the data by an increase in concentration. Figure 14 shows a similar set of curves for $H = 4470$ with various exposure times. The initial concentrations were obtained from Fig. 12.

Figure 15 shows the dark-adaptation curves of Fig. 14 plotted in the usual manner. The absolute threshold is here taken as 10×10^{-6} blondel. This is above the threshold selected in Section 7 but is a good representative value for ordinary sizes of test objects. The calculated light-adaptation curves are also shown. These curves were obtained from Fig. 12 by transferring from x to the threshold helios, using Eq. (14a). Such light-adaptation curves have never been obtained experimentally. Their most outstanding feature is the extreme rapidity with which they change when the light is turned on.

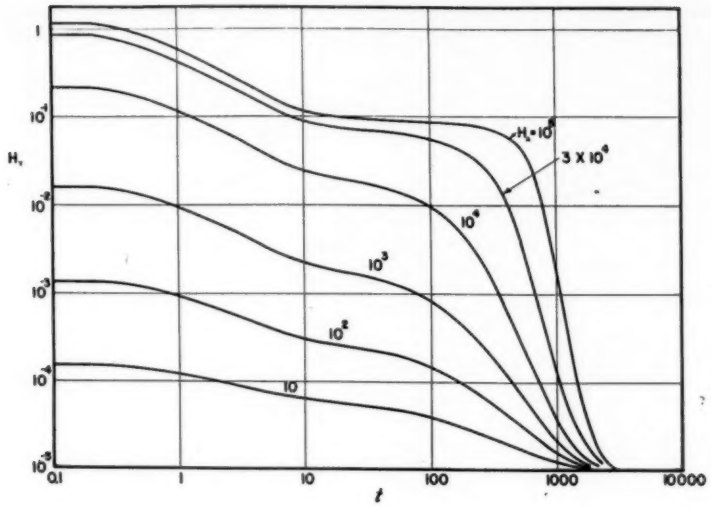


FIG. 17. Calculated dark-adaptation curves. Rod vision, natural pupil, $\tau\rightarrow\infty$.

Note the much greater slope of the light-adaptation curves compared with the dark-adaptation curves.

If curves are calculated for both rods and cones, the two sets can be combined by always selecting the lower of the two thresholds. Figure 16 gives an example. Various values of adapting helios are used and the exposure time is taken sufficiently long so that equilibrium obtains in each case. The rod curves change rather slowly and do not reach their final threshold even at 1500 sec. The cone curves, on the other hand, change quickly and come together at a constant threshold in less than 500 sec. Evidently if no special fixation is used, dark adaptation for $H_A=10^5$ will begin in the fovea and will fall rapidly along the upper cone curve until that curve intersects the rod curve at 1020 sec. Similar dark-adaptation curves are obtained for 3×10^4 and 10^4 blondels, with sharp breaks between the cone portion and the rod portion. The similarity of these curves and experimental curves can be seen by comparing Fig. 16 with the experimental curves of Hecht, Haig, and Chase.[29]

One of the peculiarities of these experimental curves[29] was the fact that at low values of adapting helios there seemed to be no cone curve. No explanation has been given of this peculiarity and there was a feeling that it might be caused by the difficulty of taking readings very near $t=0$. The calculated curves of Fig. 16, however, show that no matter how near one approaches $t=0$ there is no possibility of cone vision for $H_A=10^3$. The rod and cone curves do not intersect, and all vision at these values of H_A will be rod vision.

All the previous curves have been calculated for a 2-mm artificial pupil. But any of them can be altered to make them applicable to vision with natural pupil. The equations of Section 6 are employed. The dark-adaptation curves of Fig. 16 are replotted in Fig. 17 after correcting for artificial pupil. Figure 17 is for rod vision and natural pupil. The change in the curve for $t<100$ sec. is caused largely by the opening of the pupil, while the remainder of the variation is produced by the photochemical process. As pointed out in a paper on the transient Stiles-Crawford effect,[6] our calculations show that the photochemical process in the retina does not produce much change in the threshold at the beginning of dark adaptation. The rapid change in pupil size is of real service in lowering the threshold when such a change is badely needed.

A similar set of curves for the cones is shown in Fig. 18. The data were first calculated for a 2-mm pupil and were then corrected in the usual manner. The maximum variation in threshold

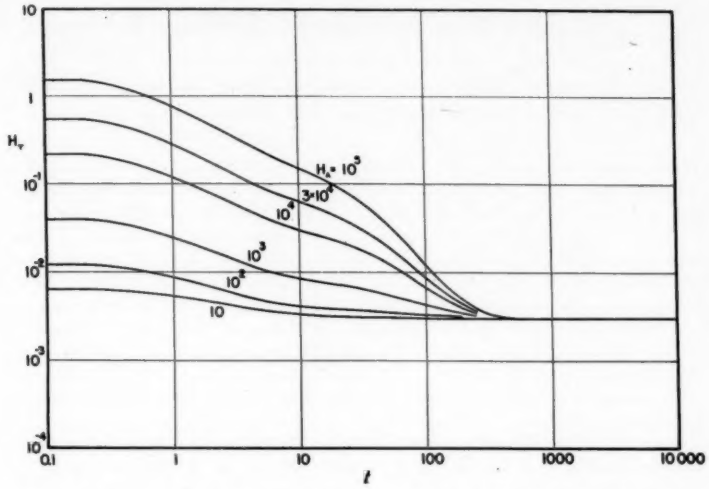


FIG. 18. Calculated dark-adaptation curves. Cone vision, natural pupil, $\tau \rightarrow \infty$.

for the cones is seen to be only about 100 to 1, while the corresponding range of H_T for the rods is over 10^4 to 1. This is also in agreement with experimental research.

11. SUMMARY

The photochemical theory of vision in its previous forms was unable to describe the recent dark-adaptation results which have been obtained with a large range of adaptation time and helios. The present study was made to see if the Hecht equations could not be modified to bring them into agreement with modern experimental research and if they could not be formulated in a more precise manner. Five postulates are stated. The rest of the paper develops the mathematical consequences of these five postulates. Solutions are obtained for the differential equation under a wide variety of conditions and the effects of area and pupil diameter are formulated.

In most cases, the predictions of the modified theory are in quantitative agreement with experiment. In other cases, the experimental data are inconsistent and it is hard to determine whether the new theory is sufficiently accurate or not. In still other cases there are as yet no experimental data by which the predictions can be checked.

The new mathematical formulation allows the calculation of an almost unbounded variety of results. It is hoped that these predictions will be a stimulus for further experimental investigations. Only by such experiments will it be possible to determine how much of this theory is good and how much will need further revision.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 35, NUMBER 1 JANUARY, 1945

Electro-Optical Properties of Colloids. II*

BERNHARD W. SAKMANN
George Eastman Research Laboratory of Physics, Massachusetts Institute of Technology, Cambridge, Massachusetts

(Received September 12, 1944)

This paper deals with investigations of the electric birefringence of colloidal solutions in alternating fields; the most detailed measurements were carried out with monodisperse bentonite sols. In general the double refraction consists of a direct and an alternating part, both of which depend on field strength and frequency of the applied field, particle size, temperature, and concentration of the solution. The direct birefringence, which can be positive or negative, can reverse its sign with a change of any of the five parameters mentioned. The measurements are interpreted by means of a relaxation theory which can explain all phenomena connected with the vibrating component in a satisfactory way. Likewise this theory is capable of explaining the variations and the reversal of the direct birefringence with changes of the frequency, field strength, particle size, and temperature. The character of the dependence of the steady birefringence on concentration which has not hitherto been observed with other colloids was first discussed by H. Mueller. The changes and reversal of sign of the direct component with concentration are accounted for by the assumption that the micelles are surrounded by a compressed water hull; the pressure of the water hull depends on particle size and concentration. Direct and alternating birefringence have different relaxation times.

INTRODUCTION

IN a previous publication[1] the experimental means of measuring electric birefringence were critically reviewed, and it was shown that compensator methods, which have been used for such investigations, are inadequate for measurements of the electro-optical effect in alternating fields. If an alternating field is applied to the plates of the Kerr cell, the birefringence consists, in general, of a direct and an alternating part:

$$n_p - n_n = a + b\cos(2\omega t - \varphi). \qquad (1)$$

n_p and n_n are the indices of refraction of light vibrating parallel and normal to the electric field of circular frequency ω, a is the direct component of the double refraction, and b is the amplitude of the alternating part which lags behind the field by an angle φ.

The electric birefringence of colloidal solutions has been the subject of many investigations.[2] Most previous workers used compensators for investigations in alternating fields without estimating the relative magnitude of b and a. Likewise, at the start of the present investigations compensators were used for the measurements, and some of the results thus obtained are reported here. Only in the course of the investigations did it become evident that such measurements give an incomplete picture. The photoelectric method was then developed and applied in all subsequent measurements; with this method the three quantities a, b, and φ of Eq. (1) were determined. The values of the direct birefringence obtained with the photoelectric method were continuously checked by compensator measurements. A similar control was not possible for measurements of the vibrating component. It is possible to measure alternating birefringence with compensators if stroboscopic illumination is used.[3] Errera, Ostveen, and Sack[4] carried out such measurements of the magneto-optical effect of colloidal solutions of iron oxide with alternating magnetic fields. For a successful application of the stroboscopic method to the present investigations serious obstacles would have to be overcome since the

* The greater part of this paper is from a dissertation submitted in partial fulfillment of the requirements for the degree of Ph.D. at Massachusetts Institute of Technology, 1941. The thesis was written under the supervision of Professor Hans Mueller.

[1] H. Mueller and B. W. Sakmann, J. Opt. Soc. Am. **32**, 309 (1942).

[2] References of previous work on the subject are given in the first article referred to above. In addition to the publications previously quoted, a good bibliography can be found in an article by W. Heller, Rev. Mod. Phys. **14**, 390 (1942). See also B. W. Sakmann, M. S. Thesis (1939) and Ph.D. Thesis (1941), Massachusetts Institute of Technology.

[3] M. Schwob and R. Lucas, Comptes rendus **194**, 1729 (1932).

[4] J. Errera, W. Oostveen, and H. Sack, Rec. Trav. Chim. **57**, 185 (1938).

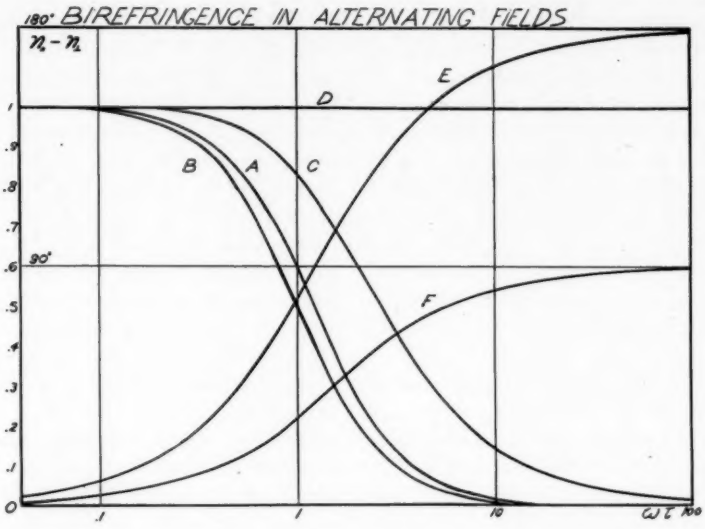


FIG. 1. Frequency dependence of direct and alternating components of double refraction and of phase angle between alternating field and vibrating birefringence for linear and quadratic effect. Curve A: vibrating component of linear effect; curve B: direct component of linear effect; curve C: vibrating component of quadratic effect; curve D: direct component of quadratic effect; curve E: phase angle for linear effect; curve F: phase angle for quadratic effect.

alternating double refraction is very small. However, it will be shown below that the results obtained with the photoelectric method are in such good agreement with the theory that it is not probable that this method leads to serious mistakes.

If the alternating component b vanishes, compensator measurements give a complete picture. Moreover, if $b \ll a$, compensators furnish accurate values of the direct birefringence, although the description of the electro-optical effect is incomplete inasmuch as the alternating part is neglected. Even if the vibrating component is comparatively weak, it can give valuable information on the structure of the investigated colloids.

As three quantities were measured instead of only one, it was not astonishing that the electro-optical effect proved to be more complex than has heretofore been assumed. On the other hand, the measurements may furnish information on size and shape of the micelles.

The first chapter of this paper outlines the scope of the investigations, gives a brief summary of the phenomenological aspects of the problem, and deals with the range of the investigated parameters.

The next section provides a critical review of the different attempts to explain the phenomena of electric double refraction.

In the third chapter it is shown how the premises of these theories have to be modified in order to be applicable to the problems under discussion. A relaxation theory which seems well adapted to the birefringence of colloids in alternating fields is presented.

The results of the measurements are discussed in the fourth chapter and are compared with the conclusions of the relaxation theory; it is shown that most of the experimental results can be explained in a satisfactory way.

The conclusions to be drawn from this work are presented in the last chapter; furthermore, the inadequacies and the direction of a possible extension of the relaxation theory are also discussed.

SCOPE OF INVESTIGATIONS

Of the colloids investigated the most extensive measurements were made with monodisperse

solutions of Wyoming bentonite.* These solutions were chosen as the subject of more detailed investigations for the following reasons.

The electro-optical effect of bentonite sols is in many respects typical of other colloidal solutions. In addition to those effects that can be found in other colloids bentonite sols exhibit a number of anomalous properties which have so far not been observed in other solutions. Furthermore, if proper precautions are taken, disturbing effects complicating the evaluation of the measurements can be made negligible in monodisperse sols.

The magnitude of the vibrating component of the birefringence of bentonite sols decreases rapidly with the frequency of the applied field. This decrease of the alternating component corresponds closely to the frequency dependence of other investigated solutions such as the colloids of vanadium pentoxide and tobacco mosaic virus. The measurements which were made with these solutions are not reproduced here.

The steady part of the birefringence can be positive or negative depending on the value of the field strength and frequency of the applied field, particle size, concentration, and temperature of the solution. Moreover, under certain conditions, the steady birefringence can reverse its sign with a change of any of these five parameters.

The voltage applied to the plates of the Kerr cell was supplied by a beat frequency oscillator whose frequency could be varied between 100 and 15,000 cycles. The 60-cycle line furnished the lowest frequency employed for the measurements. The investigations were confined to acoustical frequencies because the relaxation time of the particles of all investigated solutions was such that the vibrating birefringence decreased in the acoustical range. Furthermore, the direct birefringence of bentonite sols reverses its sign and shows its most conspicuous variations in the same frequency range. For frequencies greater than 10 kilocycles the value of the direct birefringence of bentonite sols does not change appreciably, and the value of the direct double refraction at 10 megacycles is practically identical with the one found at 10 kilocycles.

In order to avoid the disturbing influence of streaming birefringence, the value of the electric field strength has to be limited. If a voltage is applied to the plates of the Kerr cell, an electric current passes through conducting solutions. If temperature gradients are allowed to build up, the resulting convection currents give rise to streaming birefringence, and the value of the measured double refraction is the result of a superposition of streaming and electric birefringence; it would be impossible to separate the two effects.

The heat dissipated in the solution is proportional to its conductivity and increases with the square of the applied field. The disturbing influence of the heating effects can be restricted by limiting the magnitude of the electric field and by reducing the conductivity of the solution. This was achieved by using dialyzed solutions and by limiting the field on the Kerr cell to a maximum value of 200 volts/cm. Furthermore, the Kerr cell was immersed in a calorimeter which absorbed the heat generated during the measurements.

Together with the disturbing influence of temperature gradients in the solution one should mention another spurious effect which has to be guarded against in connection with measurements of the temperature dependence of the birefringence. If bentonite sols are heated to temperatures over 80°C, the structure of the colloids changes permanently, and subsequent cooling

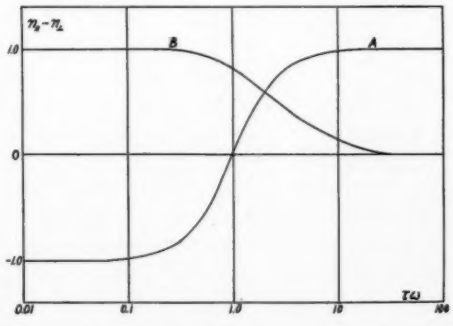


FIG. 2. Direct and alternating components of birefringence for superposition of linear and quadratic effect $b/kT : (a/kT)^2 = -\frac{1}{2}$. Curve A: direct component; curve B: alternating component.

* The investigations were made possible by Professor Hauser and Dr. Le Beau who supplied the author with monodisperse bentonite sols. The mean particle diameters mentioned below are the values determined with the supercentrifuge. See E. A. Hauser and C. C. Reed, J. Phys. Chem. **40**, 1169 (1936).

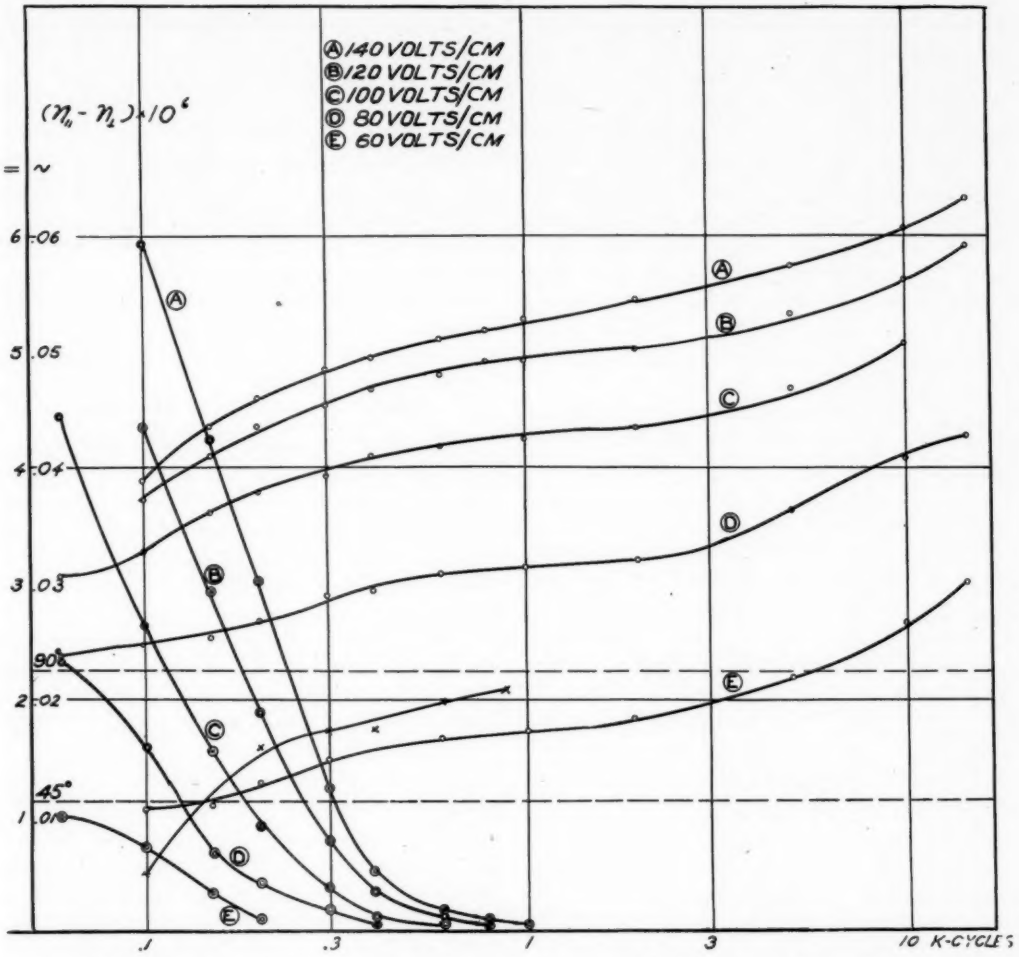


FIG. 3. Frequency dependence of direct and alternating component of birefringence of monodisperse bentonite sol of 110-mμ mean particle diameter and concentration 0.16 percent for different values of the electric field strength. Phase angle between electric field and vibrating double refraction. Direct part: ○; alternating part: ⊙; phase angle: x.

of the solutions does not restore the original value of the birefringence.

Bentonite sols whose concentrations exceed a value of 3 percent become spontaneously birefringent. In order to be entirely free from the influence of spontaneous double refraction, the concentration of all solutions discussed here was kept below 1.5 percent.

Many colloidal solutions become dichroic in an electric field. Unlike streaming birefringence caused by uneven heating of the solutions, the influence of dichroism can easily be accounted for. However, the evaluation of the experimental data is simplified considerably if the measurements are confined to those solutions whose dichroism is negligible compared with their birefringence. Bentonite sols whose mean particle diameter exceeds a value of 90 mμ show an appreciable dichroism. The dichroism of monodisperse sols whose mean particle diameter varies between 20 mμ and 90 mμ is negligible. Most of the investigations reported here were restricted to this range of particle size. The dichroism of polydisperse solutions, on the other

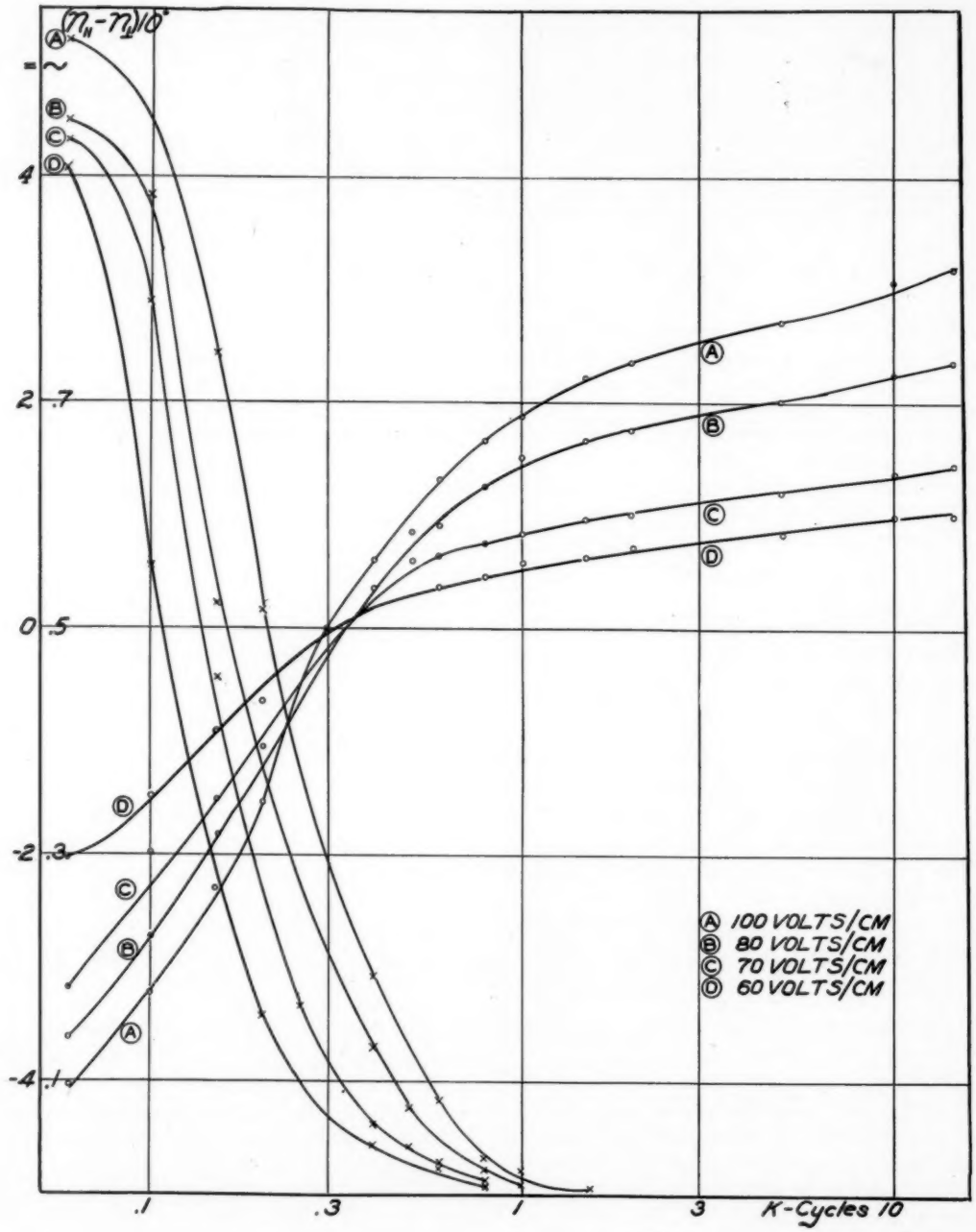


FIG. 4. Frequency dependence of direct and alternating component of birefringence of monodisperse bentonite sol of 20-mμ mean particle diameter and concentration 0.33 percent for different values of the electric field strength. Direct part: O; alternating part: x.

hand, is by no means negligible. H. Mueller and H. H. Helm[5] have shown what effects might be expected if dichroism and birefringence are of equal magnitude.

Furthermore, it is desirable to use monodisperse solutions in order to have well-defined experimental conditions. Majorana[6] and, later on, Cotton and Mouton[7] observed that the magnetic birefringence of Bravais iron reversed its sign as the magnetic field strength was increased. The latter two authors explained the reversal of sign by the assumption that different particle sizes gave rise to birefringence of opposite sign.

THEORETICAL CONSIDERATIONS

Most investigators explain the phenomenon of electric birefringence by orientation of anisotropic particles; the theory based on this view is in quantitative agreement with experimental results only for gases. Despite this fact the orientation hypothesis has been used for a qualitative interpretation of the effect in liquids and colloidal solutions. Although there are indications that the orientation theory is not entirely adequate, the measurements discussed in this paper will be interpreted on the basis of the same assumptions.

For that reason it seems advisable to review briefly the premises of Langevin's[8] orientation theory and the extensions proposed by other investigators. It will become apparent that none of these explanations is applicable to the electric birefringence of the investigated colloids and it will become necessary to formulate a theory which is better adapted to the problems under discussion.

Langevin based his theory on the following premises: (1) The oriented particles are assumed to be independent of one another. (2) The torque exerted by the electric field and the disordering influence of the Brownian motion are the only forces acting on the particles. (3) The energy of orientation of the particles is assumed to be small compared with the thermal energy of Brownian motion.

As inertia effects are neglected and the surrounding medium is assumed not to exert any forces on the particles, the birefringence follows the fluctuations of the field instantaneously. Because of assumptions (1) and (2) it is justified to assume that the distribution function of the particles is determined by Maxwell-Boltzmann's law. It is evident that a theory based on these assumptions yields good results only for gases. Liquid molecules, on the other hand, are so closely packed that the first assumption does not hold.

Peterlin and Stuart[9] pointed out that, in certain respects, the electric birefringence of colloidal solutions is analogous to the Kerr effect in gases treated by Langevin. They considered dilute colloids of uncharged dielectric particles suspended in a non-conductive medium. Inasmuch as their calculation is restricted to colloids of low concentration and to static electric fields, the first two assumptions mentioned above are correct, and Maxwell-Boltzmann's law can be applied.

In the case of the molecular Kerr effect the energy of orientation is much smaller than the thermal energy of Brownian motion even for field values near the electric breakdown strength of most liquids; the third assumption is, therefore, valid over a wide range of the electric field strength. The energy of orientation of colloidal particles, on the other hand, cannot be assumed to be small compared with the energy of Brownian motion even for comparatively small values of the electric field. Thus the third assumption mentioned above is not valid any more. Peterlin and Stuart calculated the birefringence of colloidal solutions for field values comparable to the thermal energy of Brownian motion.

In the case of the birefringence of liquids other factors have to be considered. It has long been known that the Kerr effect in liquids has a finite relaxation time; the liquid molecules do not follow fluctuations of the field instantaneously.[10] It is obvious that the assumption of independent molecules does not hold in the case of liquids.

[5] H. Mueller and H. H. Helm, J. Opt. Soc. Am. **32**, 743A (1942).
[6] Q. Majorana, Acc. dei Lincei **11**, 1° sem., 374, 463; 2° sem., 90, 139 (1902).
[7] M. Cotton and H. Mouton, Ann. Chim. Phys. **11**, 145, 289 (1907).
[8] P. Langevin, Le radium **7**, 233, 249 (1911); M. Born, Ann. d. Physik **55**, 177 (1918).
[9] A. Peterlin and H. A. Stuart, Zeits. f. Physik **112**, 129 (1939).
[10] H. Abraham and J. Lemoine. Comptes rendus **129**, 206 (1899).

Debye[11] and Tummers[12] took account of these interaction forces by the assumption that a change of orientation of the molecules is opposed by a torque T caused by the viscosity of the liquid: $T=-\rho(d\vartheta/dt)$. The friction constant ρ depends on the geometrical properties of the particles, the viscosity, and the temperature of the surrounding medium; $d\vartheta/dt$ is the angular velocity of the particles. Because of the fact that assumptions (1) and (2) do not hold any longer, there is a time lag between the orientation of the particles and the variations of the electric field, and it is no longer permissible to apply Maxwell-Boltzmann's law for the determination of the distribution function of the particles. Tummers[12] showed that if a varying torque is acting on the particles, the distribution function $f(\vartheta, t)$ must satisfy the differential equation

$$\frac{\rho}{kT}\frac{\partial f}{\partial t}=\frac{1}{\sin\vartheta}\frac{\partial}{\partial\vartheta}\left[\sin\vartheta\left(\frac{\partial f}{\partial\vartheta}-\frac{fM}{kT}\right)\right], \quad (2)$$

where k is Boltzmann's constant, T the absolute temperature, M the torque exerted by the electric field on the particles, and ϑ is the angle between the electric field and the direction of greatest dielectric polarizability of the particles.

Debye and Tummers defined the relaxation time τ and derived for nonpolar particles the expression

$$\tau=\rho/2kT. \quad (3)$$

Similarly it can be shown that for polar particles

$$\tau=\rho/6kT.$$

Values of the friction constant ρ have been determined in a few cases. For spherical particles

$$\rho=8\pi\eta r^3. \quad (4)$$

Edwardes[13] computed the frictional torque on ellipsoidal particles. For the friction constant ρ of particles of axial symmetry Gans[14] derived expressions which assume simple forms for rods and circular disks.

For rod-shaped particles the value of the friction constant depends on whether the axis of rotation coincides with the axis of symmetry or is perpendicular to the figure axis. For circular disks, on the other hand, Gans showed that the values for the friction constant ρ are identical for rotation around the axis of symmetry and perpendicular to the figure axis. Gans derived for its value the expression:

$$\rho=\frac{32}{3}\eta a^3. \quad (5)$$

The most comprehensive work on the Brownian motion of ellipsoidal particles was published by F. Perrin.[15]

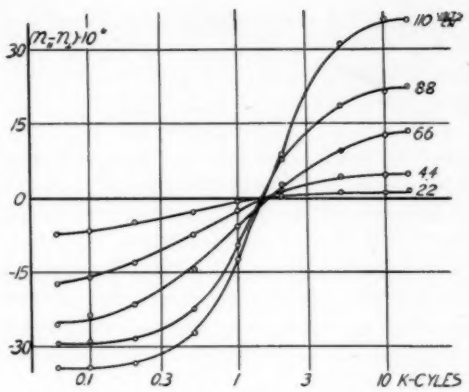


FIG. 5. Frequency dependence of direct component of birefringence of monodisperse bentonite sol of 20-mμ mean particle diameter and concentration 1.16 percent for different values of the electric field strength.

RELAXATION THEORY OF ELECTRIC BIREFRINGENCE OF COLLOIDS*

It is now possible to formulate a theory that is better adapted to conditions prevailing in colloidal solutions. As in previous theories it will be assumed that the double refraction of colloidal solutions is caused by the orientation of anisotropic micelles. For simplicity's sake the calculations will be confined to particles of axial symmetry. Most previous investigators considered orientation of polar and non-polar dielectric particles suspended in a dielectric solvent. However, it is evident that in most cases the torque exerted on the micelles is of a more complex nature. The sols which are the subject of the

[11] P. Debye, Ber. d. D. Phys. Ges. **15**, 777 (1913).
[12] J. H. Tummers, Dissertation, Utrecht (1914).
[13] D. Edwardes, Quart. J. Math. **26**, 70 (1893).
[14] R. Gans, Ann. d. Physik **86**, 652 (1928).
[15] F. Perrin, J. de phys. et rad. [5] **7**, 497 (1934); see also A. Budó, E. Fischer, and S. Miyamoto, Physik. Zeits. **40**, 337 (1939).

* Many concepts presented in this chapter were suggested to the author by Professor Mueller.

present paper consist of charged dielectric particles suspended in a conducting solvent.

Besides the torque owing to the different dielectric properties of solute and solvent one has to consider a number of other factors which give rise to a torque on the micelles. As the charged particles oscillate in the vibrating electric field, hydrodynamic forces come into play. Furthermore, one has to remember that water is a conducting medium; in addition to the hydrodynamic forces one has to consider the fact that dielectric particles suspended in a conducting medium are acted upon by a torque. Moreover, the ion cloud surrounding the charged particles is distorted in the alternating electric field, and it seems probable that an asymmetric arrangement of the ions around the micelles gives rise to an additional torque.

At the present time it does not seem possible to determine the relative importance of the different factors involved. For anisotropic dielectric polar particles the torque has two components: The torque caused by the permanent dipole which is directly proportional to the field strength

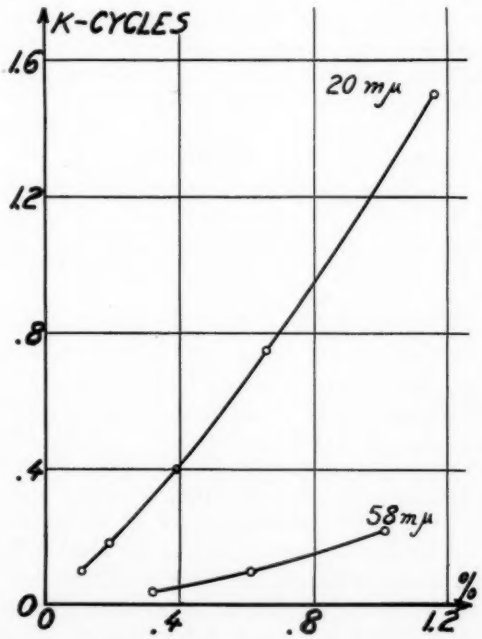


FIG. 6. Dependence of reversal frequency on concentration for two monodisperse bentonite sols of 20-mμ and 58-mμ mean particle diameter.

and the moment caused by the anisotropy of the micelles which is proportional to the square of the field. The torque has the form

$$M = aE_0 \cos\vartheta \cos\omega t + bE_0^2 \sin\vartheta \cos\vartheta \cos^2\omega t. \quad (6)$$

In the dielectric case, a is equal to the dipole moment and b equals half the value of the dielectric anisotropy. Although it is evident from the discussion above that conditions prevalent in colloidal solutions are much more complicated than in the dielectric case, it will be assumed in the following that the torque on the particles has the form of Eq. (6). Such an assumption can be justified only on empirical grounds, and one has to see whether the results of the computations based on this assumption are in agreement with the measurements.

It must be admitted that the physical significance of a and b is not evident, and it is not excluded that the two factors are functions of the frequency ω of the applied field. Also it is possible that the electric field and torque on the particles are out of phase; for simplicity's sake this case will not be considered at the present time. It is assumed that the distribution function of the particles follows the differential equation (2). Furthermore, it is evident that the condition has to be satisfied:

$$\int_0^{2\pi}\int_0^{\pi} f(\vartheta)\sin\vartheta d\vartheta d\varphi = N, \quad (7)$$

where N is the number of particles per unit volume. Condition (7) can be satisfied if $f(\vartheta)$ is represented as a sum of spherical harmonics. A solution of differential equation (2) for a torque of the form (6) can best be found by considering the distribution function for the quasistatic case.

If terms beyond the second power in E are neglected, the distribution function $f(\vartheta, t)$ for the static case is according to Maxwell-Boltzmann:

$$f(\vartheta, t) = \frac{N}{4\pi}\left[1 - \frac{aE_0}{kT}\cos\omega t\cos\vartheta + \frac{E_0^2}{2}\cos^2\omega t\,(\cos^2\vartheta - \tfrac{1}{3})\left\{\frac{b}{kT} + \left(\frac{a}{kT}\right)^2\right\}\right]. \quad (8)$$

Analogously to solution (8) one may try a solu-

tion of the form:

$$f(\vartheta, t)=\frac{N}{4\pi}\left[1-\frac{aE_0}{kT}\cos\vartheta p(t)\right.$$
$$\left.+\frac{E_0^2}{2}(\cos^2\vartheta-\tfrac{1}{3})\left\{\frac{b}{kT}q(t)+\left(\frac{a}{kT}\right)^2 r(t)\right\}\right]. \quad (9)$$

The time functions $p(t)$, $q(t)$, and $r(t)$ are found by inserting Eq. (9) into the differential equation (2). By comparing terms in $\cos\vartheta$ and $(\cos^2\vartheta-\frac{1}{3})$ one finds for the three time functions the following differential equations:

$$p(t)+\tau\frac{dp}{dt}=-\cos\omega t, \quad (10)$$

$$q(t)+\frac{\tau}{3}\frac{dq}{dt}=-\tfrac{1}{2}(1+\cos 2\omega t), \quad (11)$$

$$r(t)+\frac{\tau}{3}\frac{dr}{dt}=\cos\varphi\cos\omega t\cos(\omega t-\varphi). \quad (12)$$

The solutions of Eqs. (10) to (12) are found to be:

$$p(t)=-\cos\varphi\cos(\omega t-\varphi), \quad (13)$$

$$q(t)=\tfrac{1}{2}[1+\cos\psi\cos(2\omega t-\psi)], \quad (14)$$

$$r(t)=\tfrac{1}{2}[\cos^2\varphi+\cos\psi\cos\varphi \times\cos(2\omega t-\varphi-\psi)], \quad (15)$$

where

$$\tan\varphi=\tau\omega, \quad (16)$$

$$\tan\psi=\tfrac{2}{3}\tau\omega. \quad (17)$$

$p(t)$ plays a role in Debye's[11] theory of anomalous dispersion of the dielectric constant, and the function $q(t)$ was derived by Tummers[12] in his theory of the electric birefringence of non-polar liquids. The distribution function $f(\vartheta, t)$ is therefore

$$f(\vartheta, t)=\frac{N}{4\pi}\left[1+\frac{aE_0}{kT}\cos\vartheta p(t)\right.$$
$$\left.+\frac{E_0^2}{2}(\cos^2\vartheta-\tfrac{1}{3})\left\{\frac{b}{kT}q(t)+\left(\frac{a}{kT}\right)^2 r(t)\right\}\right]. \quad (18)$$

If α_1 and α_2 are the two values of the optical polarizability, the birefringence is found to be:

$$n_p-n_n=\frac{3\pi}{n_0}(\alpha_1-\alpha_2) \times\int_0^\pi f(\vartheta, t)(\cos^2\vartheta-\tfrac{1}{3})\sin\vartheta d\vartheta. \quad (19)$$

From Eq. (19) it is evident that the term directly proportional to the electric field does not contribute anything to the value of the birefringence, and, for the evaluation of the integral, only the terms proportional to the square of the field have to be considered. The final result of the integration is:

$$n_p-n_n=\frac{2}{15}N\frac{\alpha_1-\alpha_2}{n_0}E_0^2$$
$$\times\left\{\frac{b}{2kT}[1+\cos\psi\cos(2\omega t-\psi)]\right.$$
$$\left.+\frac{1}{2}\left(\frac{a}{kT}\right)^2[\cos^2\varphi+\cos\varphi\cos\psi \times\cos(2\omega t-\varphi-\psi)]\right\}. \quad (20)$$

The first and second terms of Eq. (20) arise from torques which are proportional to E_0^2 and E_0, respectively. In the following the birefringence originating from torques proportional to E_0 and E_0^2 will be referred to as linear and quadratic effect, respectively. However, it is necessary to keep in mind that quadratic as well as linear torques give rise to birefringence which is proportional to E_0^2. In the dielectric case polar orientation corresponds to the linear effect, whereas the orientation of non-polar particles corresponds to the quadratic effect.

As in the case of the molecular Kerr effect of gases treated by Langevin and Born[6] the temperature dependence is different for linear and quadratic effect, the former being proportional to $1/T$, the latter to $1/T^2$.

As the functions $q(t)$ and $r(t)$ differ in several respects, it should be possible to decide whether the birefringence is caused by a linear or a quadratic torque on the micelles or by a combination of both. Equations (14) and (15) show that both functions have a steady and a vibrating part. In both cases the alternating component decreases with the frequency of the applied field in a similar fashion. However, the character of the direct part of the two functions is very different. While the steady part of $q(t)$ is independent of frequency, the direct component of $r(t)$ decreases with the frequency in a way similar to the alternating part.

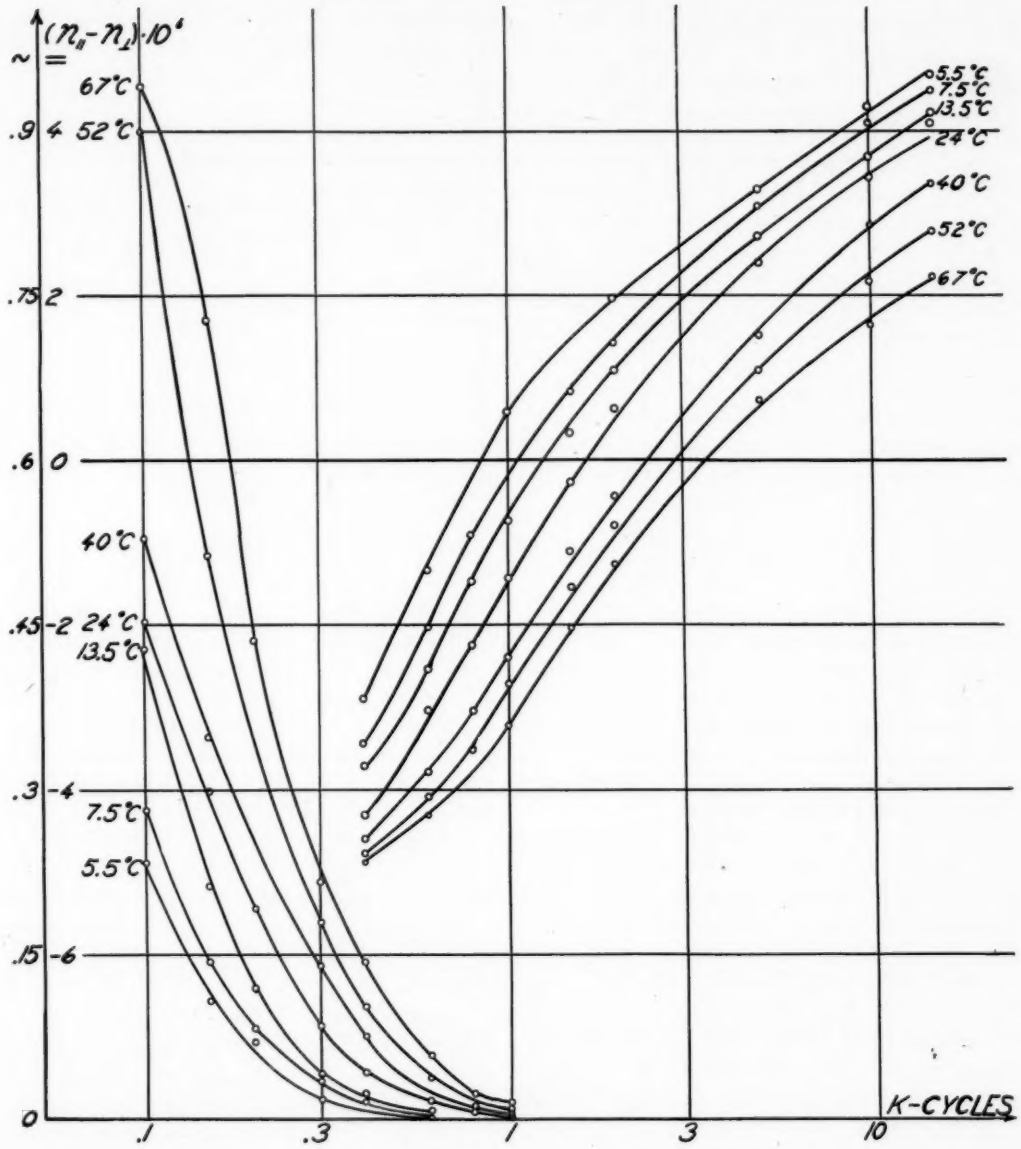


FIG. 7. Frequency dependence of direct and alternating component of the birefringence of a monodisperse bentonite sol of 20-mμ mean particle diameter and concentration 1.16 percent for constant field strength of 40 volts/cm and different values of the temperature of the solution.

In Fig. 1 the two functions $q(t)$ and $r(t)$ are plotted *versus* $\omega\tau$. Curves A and C represent the frequency dependence of the alternating parts of $r(t)$ and $q(t)$, respectively. A comparison of the two curves shows that the frequency dependence of the vibrating parts is similar, and that it would be difficult to distinguish between the two cases on the basis of measurements of the alternating birefringence alone. However, the frequency dependence of the steady part of the two

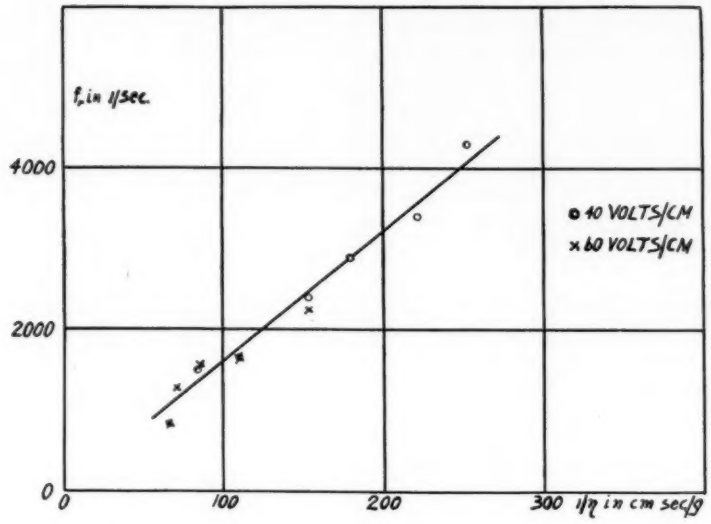


FIG. 8. Reversal frequency in cycles per sec. *versus* inverse of viscosity in cm sec./g for monodisperse bentonite sol of 20-mμ mean particle diameter and concentration 1.16 percent.

functions is very different. Curves B and D show the frequency dependence of the direct part of $r(t)$ and $q(t)$, respectively. The steady part of $r(t)$ decreases in a manner similar to the one observed for the vibrating components; the direct part of $q(t)$, on the other hand, is constant over the whole frequency range.

A further difference between linear and quadratic effect can be found in the frequency dependence of the phase angle between vibrating birefringence and applied field. Curves E and F show the phase angle dependence of the two functions. The plot shows that for high frequencies the phase angles approach asymptotically values of 90° and 180° for $q(t)$ and $r(t)$, respectively. It is evident that phase angle measurements may provide a convenient means of distinguishing between linear and quadratic effect. However, if one is dealing with a superposition of the two effects such measurements are more difficult to interpret.

In the first article it was shown that the phase angle between the electric field on the Kerr cell and the resulting birefringence was determined from the Lissajous figure observed on the screen of the cathode-ray oscilloscope. This means of measuring phase angles is not able to distinguish between a leading and a lagging phase shift; in other words, this method cannot distinguish between a phase angle φ and $360-\varphi$. It will be shown that this fact is important for the interpretation of the measurements discussed below.

In general one might deal with either of the two effects or a superposition of both. The birefringence caused by linear and quadratic torque may have the same sign, or the two torques may give rise to double refraction of opposite sign. In the first case it is easy to see how a superposition of the two effects will influence the birefringence. The second case in which the two effects are of opposite sign is of greater interest and will be discussed in more detail.

Because of its importance to the present problem the discussion will be limited to the case in which the linear torque gives rise to negative birefringence and the quadratic torque causes a positive double refraction. For the dielectric case, such a situation may arise from superposition of polar and nonpolar orientation causing negative and positive double refraction. If the dipole moment does not coincide with the direction of greatest optical polarizability of the particles, polar orientation may give rise to negative birefringence.

One special case will be discussed in more detail for which the constants a and b have the ratio:

$$b/kT:(a/kT)^2=-\tfrac{1}{2}. \quad (21)$$

In the discussion of the experimental results it will become apparent that the frequency dependence of the direct birefringence suggests this particular choice of the constants a and b. Equation (20) shows that for this ratio of the constants a and b the birefringence assumes the following value:

$$n_p-n_n=C[(1-2\cos^2\varphi) -\cos\psi\cos(2\omega t-2\varphi-\psi)], \quad (22)$$

where the angles φ and ψ are defined according to Eqs. (16) and (17), and the constant C has the value:

$$C=\frac{1}{15}N\frac{\alpha_1-\alpha_2}{n_0}E_0^2\frac{b}{kT}.$$

Direct part A and alternating part B are plotted in Fig. 2. It is evident from Eq. (22) that the phase angle between electric field and vibrating birefringence approaches asymptotically a value of 270° for high frequencies.

EXPERIMENTAL RESULTS

In this chapter the measurements are discussed, and the results are compared with the conclusions of the relaxation theory. The measurements are reported in the order of the investigations of the different parameters; the frequency dependence of the two components of the double refraction is presented first.

(a) Frequency

Figure 3 shows the birefringence of a large particle sol of 110-mμ mean particle diameter and a concentration of 0.16 percent. The direct double refraction and the r.m.s. value of the alternating birefringence are plotted along the ordinate, the frequency of the field applied to the Kerr cell is plotted along the abscissa on a logarithmic scale; the field strength is the independent parameter. As the maximum value of the alternating effect is smaller than the direct birefringence by two orders of magnitude, the scale for the vibrating component is chosen 100 times greater than that for the direct effect. The decrease of the alternating birefringence with increasing frequency seems to be in good agreement with the results of the relaxation theory as is shown by a comparison of curves A and C of Fig. 1 with the experimental curves of Fig. 3. It is evident from the curves of Fig. 1 that the vibrating double refraction reaches a constant value for frequencies which are small compared with the relaxation frequency of the micelles. Because the frequencies available for our measurements were not low enough, this maximum value of the alternating part was not reached. Furthermore, the phase angle between alternating birefringence and applied field is plotted in Fig. 3. The plot shows that a value of 90° is approached asymptotically for high frequencies.

In this connection it is necessary to keep in mind that the phase angles were determined from the Lissajous figure observed on the screen of a cathode-ray oscilloscope. It was mentioned in the first article that the shape of the Lissajous figure is invariant toward a change of sign of the phase angle. A phase lag of 90° gives rise to the same Lissajous figure as a leading phase shift of 90° which, in its turn, is equivalent to a phase lag of 270°. The fact that a phase shift of 90° (or 270°) is approached for high frequencies is in accord with Eq. (22).

Figure 4 shows the frequency dependence of a small particle sol of 20-mμ mean particle diameter and a concentration of 0.33 percent. For this solution the maximum value of the direct birefringence is three times as great as that of the alternating component. For this reason a value of 1:10 was chosen for the ratio of the scales for direct and alternating effect instead of the ratio 1:100 of Fig. 3. Comparison of Figs. 3 and 4 makes it evident that for a given frequency and field strength the vibrating birefringence of the 20-mμ particle sol is between thirty and sixty times as great as that of the 110-mμ particle sol. This difference cannot be accounted for by the fact that the concentration of the small particle sol is greater. The reason for the increase of the alternating component has to be sought for in the difference in relaxation times of the two solutions as shown by Eqs. (3)–(5) and (20).

From Fig. 4 it is evident that the direct birefringence consists of a positive and a negative branch of about equal magnitude. The double

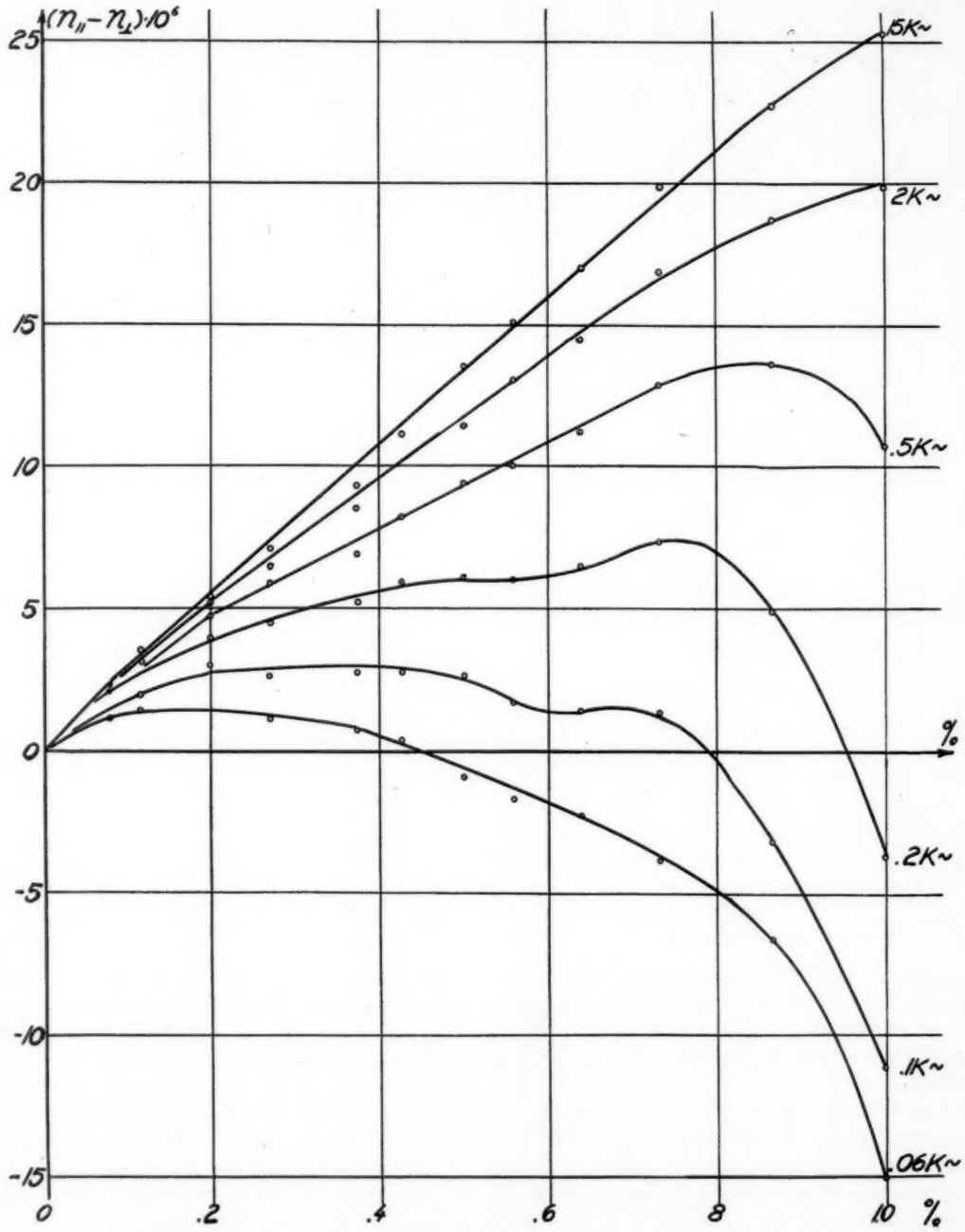


FIG. 9. Dependence of direct component of birefringence on concentration for monodisperse bentonite sol of 50-mμ mean particle diameter and for constant field strength of 85 volts/cm.

refraction is negative for frequencies below 300 cycles and positive for higher frequencies.

Figure 5 shows the frequency dependence of the direct birefringence for a solution of the same particle size of 20-mμ mean diameter, but a higher concentration of 1.16 percent. As these measurements were carried out with a Babinet compensator, no values were obtained for the alternating component. The experimental curves of Fig. 5 have exactly the same character as the theoretical curve *A* of Fig. 2. In both cases the birefringence approaches asymptotically a constant negative value for frequencies smaller than the relaxation frequency of the micelles. For high frequencies, on the other hand, the birefringence approaches a constant positive value. In both cases the variation of the direct double refraction extends over a frequency range of two octaves. Figure 4 shows that the reversal frequency has a value of 300 cycles/sec. for a concentration of 0.33 percent while the birefringence of the sol of 1.16 percent reverses its sign at a frequency of 1500 cycles/sec. From this fact it is evident that the reversal frequency depends on concentration.

This connection between reversal frequency and concentration is illustrated in Fig. 6 for two values of the mean particle diameter, 20 mμ and 58 mμ, respectively. From this plot it is evident that the increase of reversal frequency with concentration is approximately linear. Furthermore, it is interesting to note that for sols of small particle size the birefringence reverses its sign at higher frequencies than for large particle sols.

(b) Temperature

The temperature dependence of the two components of the birefringence was investigated next as it was hoped that such measurements might give information on the origin of the two effects.

The results of these measurements are shown in Fig. 7. The two components of the birefringence are plotted against the frequency of the applied field for different values of the temperature of the solution. As in previous diagrams the frequency is plotted along the abscissa on a logarithmic scale, the difference of the indices of refraction is plotted along the ordinate. For all points the field strength was kept constant at a value of 40 volts/cm. The solution of 1.16 percent concentration and 20-mμ mean particle diameter is identical with the one whose compensator measurements were shown in Fig. 5.

The investigations of the temperature dependence of Fig. 7 were made a year after the Babinet compensator measurements of Fig. 5. It is therefore not surprising that, for the same values of field strength and frequency, the values of the birefringence of the two figures show discrepancies which have to be attributed to aging of the solution.

At the time the measurements of Fig. 7 were made the room temperature was 32°C. Only two values of the temperature, 24°C and 40°C, could be kept constant throughout the measurements. All the other temperature values were changing during the measurements because of the temperature difference between the calorimeter and its surrounding. For the lowest and highest temperatures the maximum variations amounted to 1.5°C. The variations were smaller for the intermediate temperature values.

On the basis of the temperature measurements it should be possible to decide whether the torque on the micelles which causes the alternating birefringence is linear or quadratic. The double refraction caused by linear and quadratic torque differs in several respects. Equation (20) of the preceding chapter shows that the alternating birefringence which is caused by a quadratic torque has the form:

$$(n_p-n_n)_\sim = \frac{1}{15}N\frac{\alpha_1-\alpha_2}{n_0}E_0^2\frac{b}{kT} \times\left(1+\frac{4}{9}\omega^2\tau^2\right)^{-\frac{1}{2}} \cos(2\omega t-\psi), \quad (23)$$

where

$$\tan\psi=\tfrac{2}{3}\omega\tau.$$

The vibrating double refraction caused by a linear torque is

$$(n_p-n_n)_\sim = \frac{1}{15}N\frac{\alpha_1-\alpha_2}{n_0}E_0^2\left(\frac{a}{kT}\right)^2 \times\left(1+\frac{13}{9}\omega^2\tau^2+\frac{4}{9}\omega^4\tau^4\right)^{-\frac{1}{2}} \cos(2\omega t-\varphi-\psi), \quad (24)$$

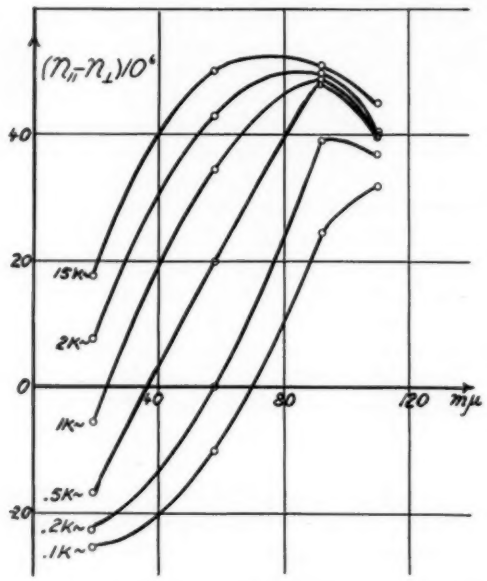


FIG. 10. Dependence of direct component of birefringence on mean particle diameter in mμ for field strength of 86 volts/cm and concentration of 1.0 percent.

where

$$\tan \varphi = \omega\tau.$$

The two equations show that the amplitudes of linear and quadratic effect are proportional to $1/T^2$ and $1/T$, respectively. This fact, however, is of only secondary importance as the temperature dependence of the relaxation time has a far greater influence. Equations (3) to (5) show that the relaxation time is directly proportional to the viscosity of the solvent which in its turn changes rapidly with temperature. Inasmuch as the dimensions of the colloidal particles are large compared with the molecular dimensions of the solvent, it seems permissible to use the macroscopic viscosity. The temperature dependence of the viscosity of water is well known, and it is possible to compute the temperature dependence of the alternating birefringence. A comparison of the experimental data with the results of the computation makes it possible to decide whether one is dealing with a linear or a quadratic effect.

The following discussion is restricted to frequencies which are large compared with the relaxation frequency of the micelles, i.e., $\tau\omega \gg 1$. For these frequencies one has for the quadratic effect:

$$(n_p - n_n)_{\sim} \omega\eta T = f_1, \tag{25}$$

and for the linear effect,

$$(n_p - n_n)_{\sim} \omega^2\eta^2 T^2 = f_2. \tag{26}$$

The products (25) and (26) were determined for the data of Fig. 7 with the following results.

For 65 different values of the birefringence the product (26) was determined, and its mean value was found to be 1.43. For all values of the product $0.65 \leq f_2 \leq 2.15$. The mean quadratic error of a single value of the product was found to be $[\Sigma\epsilon^2(n-1)]^{\frac{1}{2}} = 0.044$, where n is the number of measurements and ϵ is the difference from the arithmetic mean value. The frequency ω and the viscosity η enter as the square in Eq. (26). The error of the latter quantity is considerable because of the temperature variations during the measurements. Within the accuracy of the measurements f_2 can, therefore, be considered a constant.

For the same set of experimental data the product f_1, on the other hand, varied by more than a factor 500 between a minimum value of 1.70×10^{-5} and a maximum value 895×10^{-5}.

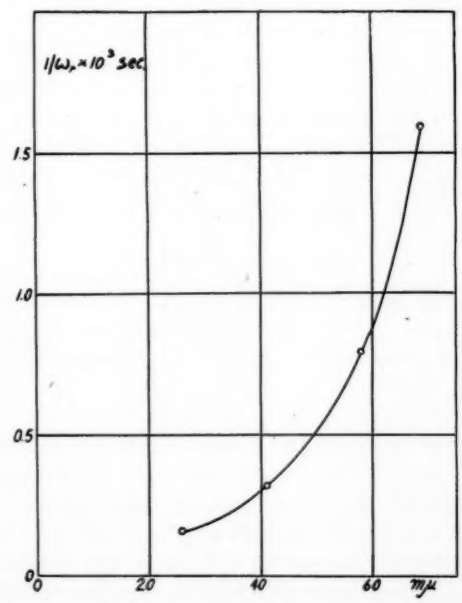


FIG. 11. Inverse of reversal frequency of direct birefringence *versus* mean particle diameter in $m\mu$ for concentration of 1.0 percent and field strength of 86 volts/cm.

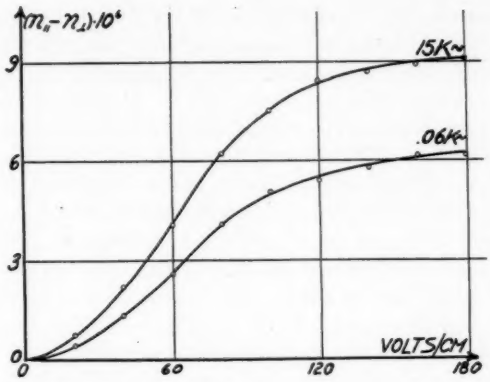


FIG. 12. Dependence of direct birefringence on field strength for monodisperse bentonite sol of 110-mμ mean particle diameter and concentration 0.14 percent.

The experimental data which were gained for an electric field strength of 60 volts/cm are not reproduced here. Computation of the products (25) and (26) gave similar results leading to the same conclusion that the torque causing the vibrating birefringence is directly proportional to the field strength.

The measurements of Fig. 7 show that the reversal frequency depends on the temperature of the solution. This is not surprising as the condition for a change of sign of the birefringence is $\omega\tau=1$. It is, therefore, to be expected that the reversal frequency is inversely proportional to the viscosity of the solvent. In Fig. 8 the reversal frequency is plotted against the inverse of the viscosity of water for the two values of the field strength of 40 and 60 volts/cm. If allowance is made for the limited accuracy of the temperature measurements, the points of Fig. 8 can well be represented by a straight line. The dimensions of the micelles can be determined from the slope of the line. If the suspended particles are assumed to be disk-shaped, Eqs. (4) and the relation $\omega\tau=1$ furnish the following expression for the dependence of the reversal frequency n on the viscosity of the solvent:

$$n=\frac{3kT}{32\pi a^3}\frac{1}{\eta}, \qquad (27)$$

where a is the radius of the disk, k Boltzmann's constant, and T the absolute temperature. From the relation (27) and the slope of the straight line of Fig. 8 one finds for the radius of the disk-shaped micelles the value $a=43$ mμ.

(c) Concentration

Comparison of Figs. 4 and 5 shows that the positive as well as the negative branch of the direct birefringence depends on the concentration of the solution, and one might expect that measurements of the concentration dependence might give further information on the structure of the colloids. Such measurements were carried out with a solution of 50-mμ mean particle diameter.

Figure 9 shows the results of these measurements which were carried out for a constant value of the electric field strength of 85 volts/cm. The frequency of the field is the parameter. It is evident from Fig. 9 that the character of the concentration dependence of the direct birefringence varies strongly with the value of the frequency of the applied field.

For the highest frequency of 15 kilocycles the birefringence is directly proportional to the concentration. Such a linear relationship is in accord with the relaxation theory of the preceding chapter; Eq. (20) shows that the birefringence is proportional to the number of particles per unit volume. Likewise for lower frequencies between 10 kilocycles and 500 cycles and concentrations smaller than 0.8 percent the increase of the birefringence with concentration is approximately linear. Deviations from the

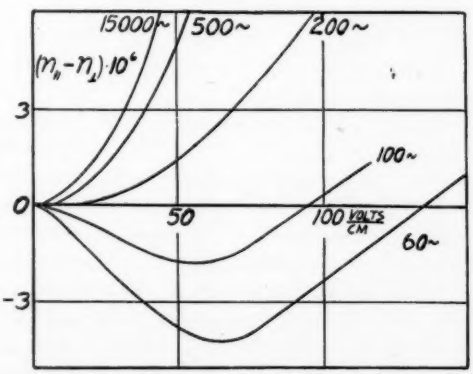


FIG. 13. Dependence of direct birefringence on field strength for monodisperse bentonite sol of 58-mμ mean particle diameter and concentration 0.77 percent.

linear relationship are perceptible for the highest concentrations only.

For low frequencies, however, the concentration dependence has an entirely different character: The positive double refraction increases for small concentrations, reaches a maximum, decreases, and finally assumes negative values; the negative birefringence increases rapidly with concentration.

The measurements of the concentration dependence of Fig. 9 are in accord with the previously discussed frequency dependence. For concentrations higher than 0.5 percent a cross section of the assembly of curves perpendicular to the abscissa gives rise to a frequency dependence as shown in Figs. 4 and 5. For small concentrations, on the other hand, the frequency dependence is similar to the one depicted in Fig. 3.

The reversal frequency is inversely proportional to the relaxation time τ of the particles, and it was shown above that the observed shift of the reversal frequency with particle size and temperature can be satisfactorily explained. However, a change of reversal frequency with concentration as shown in Fig. 9 and, more explicitly, in Fig. 6, cannot be explained in such a simple way, and it becomes necessary to look for a different explanation.

Such a shift of the reversal frequency could be explained by a change of particle size with concentration. It is well known that the micelles of hydrophilic colloids swell if gels are diluted. Hydration of the micelles would cause an increase of the relaxation time and a decrease of the reversal frequency. An increase of the reversal frequency with concentration as shown in Fig. 6 could thus be accounted for.

However, it is not probable that hydration plays an important part for the concentration range of the present measurements, and it seems advisable to look for a different interpretation of the experimental data. It might well be that the key to an understanding of these effects can be found in the measurements of Hauser and Le Beau[16] who carried out density measurements of the dissolved bentonite particles of different monodisperse solutions of various concentrations. They found that the "apparent specific gravity" of the colloidal particles depended on concentration and, for a given concentration, on the size of the micelles.

From dry weight determinations they obtained values for the weight of the dissolved bentonite and of the water of the solution. From the difference of the volume of the solution and that of the water they determined the volume of the suspended particles; they called the ratio of weight and volume of the suspended micelles their "apparent specific gravity." For concentrations below 2 percent the apparent specific gravity varied strongly with the concentration of the solution; the most conspicuous variations occurred in the concentration range between 0.1 and 1.0 percent. Hauser and Le Beau suggested that the observed variations of the apparent specific gravity are due to the presence of a highly compressed water hull surrounding the micelles. If the pressure in the water hull is so high that the specific gravity of water is increased appreciably, it is to be expected that the viscosity of the solvent and, therefore, the relaxation time of the micelles is increased likewise. One could thus explain the observed shift of the reversal frequency with the concentration of the solutions.

(d) Particle Size

In Fig. 10 the direct part of the double refraction is plotted against the mean particle diameter determined by the supercentrifuge. For all measured points of the diagram the concentration of 1.0 percent and the field strength of 86 volts/cm were kept constant.

The measurements of Fig. 10 are in line with results discussed previously. A cross section of the assembly of curves parallel to the ordinate for particle diameters above 80 mμ gives rise to a frequency dependence of the type shown in Fig. 3. For particle diameters smaller than 60 mμ a cross section results in a frequency dependence similar to the one shown in Fig. 5.

For the highest frequencies the birefringence is positive for all particle sizes and increases with the mean particle diameter. Such a dependence of the birefringence on particle size is to be explained by an increase of the optical anisotropy

[16] E. A. Hauser and D. S. Le Beau, J. Phys. Chem. **42**, 1031 (1938).

$\alpha_1 - \alpha_2$ of Eq. (20). Figure 10 shows further that the birefringence decreases slightly for the largest particles whose mean diameter exceeds a value of 80 mμ.

While it is to be expected that the birefringence should increase with particle size, it is difficult to see why the double refraction decreased for the largest particles. This decrease of the birefringence might be caused by the dichroism of the solutions which was not negligible for large particle sols. The measurements, however, were evaluated with the assumption that the birefringence of the sols was much larger than their dichroism.

It is evident from Fig. 10 that the reversal frequency increases with decreasing particle size. An increase of the mean particle diameter by about a factor 3 from 25 mμ to 70 mμ corresponds to a shift of the reversal frequency by a factor 10 from 100 to 1000 cycles/sec. The magnitude of the shift of the reversal frequency is in fairly good agreement with the change of relaxation time to be expected from Eqs. (3) to (5).

The relaxation time τ of the particles equals the inverse of the reversal frequency ω_r. From the values of Fig. 10 the relaxation times were determined for the different particle sizes; the results are plotted in Fig. 11. According to Eqs. (3) and (4) it is to be expected that the relaxation time is proportional to the cube of the diameter of the micelles. If the relaxation time is plotted against the cube of the mean particle diameter the measured points are very nearly on a straight line.

(e) Field Strength

The character of the dependence of the direct birefringence on field strength varies with the value of the particle size of the micelles and the concentration of the solution. Two typical examples of the field dependence of the direct double refraction are shown in Figs. 12 and 13. The first diagram is characteristic of a large particle sol of low concentration while the second diagram is typical for solutions of medium-sized and small particles of high concentration. In both cases the frequency of the applied field is the parameter. The field dependence of large particle sols, which is the simpler of the two cases, is discussed first.

Figure 12 shows the field dependence of the steady birefringence for a monodisperse bentonite sol of concentration 0.14 percent and 110-mμ mean particle diameter. The field dependence is shown for only two frequencies of 60 cycles and 15 kilocycles; the curves which correspond to intermediate frequency values have the same character and lie in between the two plotted curves. Analogously to the electric birefringence of gases discussed by Langevin the double refraction increases quadratically for small values of the electric field. For higher fields the increase is less rapid than the one corresponding to a quadratic relationship; finally, for the highest fields attainable, the double refraction reaches a saturation value.

Such a behavior of the direct birefringence is in good agreement with the orientation theory presented above. Peterlin and Stuart[9] discussed the saturation effect of the double refraction of colloidal solutions for steady fields. For small fields for which the energy of orientation is less than the thermal energy of Brownian motion the birefringence increases quadratically with the applied field. For strong fields for which the energy of orientation is greater than that of the Brownian motion all particles are lined up, and the birefringence reaches a saturation value.

In alternating fields similar conditions prevail for frequencies greater than the relaxation frequency of the micelles. For these high frequencies the quadratic torque on the colloidal particles predominates, thus giving rise to a positive birefringence. The fact that the saturation value of the birefringence for 60 cycles is smaller than that for 15 kilocycles might indicate that the constant b of Eq. (6) is frequency dependent.

For high frequencies the field dependence of more concentrated bentonite solutions of small and medium particle size has the same character as the curves of Fig. 12. For low frequencies, however, the observed field dependence is of a more complex nature. Figure 13 shows a typical set of Babinet compensator measurements for a sol of 58-mμ mean particle diameter and a concentration of 0.77 percent. As the measurements were made with a compensator, no values were obtained for the alternating part of the birefringence. The presence of the vibrating com-

ponent caused a broadening of the interference fringes.

At low frequencies the double refraction is negative, increasing quadratically with the field strength. If the field strength is further increased, the birefringence has a maximum in the negative region, vanishes, and assumes positive values finally reaching saturation for the highest fields attainable.

The relaxation time of these small particle sols is considerably smaller than the one of the large particle sol of Fig. 12. High frequencies between 200 cycles and 15 kilocycles are well above the relaxation frequency of the micelles, and the same conditions prevail as in the solution of Fig. 12, thus giving rise to a similar field dependence. For frequencies below the relaxation frequency, however, the linear torque comes into play causing a negative birefringence. For higher field values the quadratic torque predominates, and the double refraction is positive.

CONCLUSION

It was shown above that many aspects of the electro-optical effect of bentonite sols can be explained in a satisfactory way by orientation of the micelles in the electric field if it is assumed that any change of orientation is opposed by a torque exerted on the micelles by the viscous solvent. Furthermore, it has to be assumed that the torque caused by the electric field has two components one of which is directly proportional to the field while the other increases with the square of the field.

With these assumptions it is possible to explain quantitatively the frequency dependence of the direct and alternating birefringence. For some solutions the steady part is negative for low frequencies and reaches positive values for higher frequencies.

Likewise the change of the birefringence with field strength can be accounted for; under certain conditions the double refraction is negative for small fields, reverses its sign, and reaches a saturation value for the highest fields.

Moreover, the temperature dependence of both components of the double refraction strongly supports the orientation hypothesis. The rapid increase of the amplitude of the vibrating part as well as the shift of the reversal frequency of the direct component with temperature is in accord with the proposed theory.

Furthermore, it can be shown that the dependence of the two effects on particle size is in agreement with the proposed hypotheses. The amplitude of the vibrating component increases rapidly with decreasing particle size. Likewise the change of the direct birefringence with particle size and the dependence of the reversal frequency on the mean particle diameter are in accord with the theory presented.

It was further attempted to account for the concentration dependence of the direct birefringence by the presence of a highly compressed water hull surrounding the micelles, an assumption which is suggested by the results of other investigators. In order to be able to explain the experimental data, it is necessary to assume that the pressure in the water hull depends on concentration; this assumption seems to be supported by the measurements of Hauser and Le Beau.[16]

While the relaxation theory seems to be well adapted to the problem under discussion, it must be admitted that certain aspects of the electro-optical effect evade a satisfactory interpretation. The remainder of this chapter deals with those phenomena that cannot be interpreted by the proposed theory.

A comparison of the experimental curves of Figs. 4 and 5 with the theoretical curves of Fig. 2 shows quantitative agreement for the frequency dependence of the direct part of the birefringence. According to the relaxation theory the strongest decrease of the vibrating double refraction should occur near the reversal frequency. However, Fig. 4 shows that this requirement is not met by the result of the measurements. A possible explanation of this discrepancy may be found in the fact that, in the theoretical derivation, it was assumed that the torque on the particles is in phase with the field.

A more serious objection to the proposed theory can be found in the fact that direct and alternating double refraction have different relaxation times. The relaxation time of the direct effect is considerably greater than that of the vibrating birefringence. This could be demonstrated in the following way.

In the description of the photoelectric method

presented in the preceding publication it was shown that the intensity of the light beam incident on the photo-cell can be approximated by the equation

$$i=i_0\left[\sin^2\frac{A}{2}+\frac{B}{2}\sin A\cos(2\omega t-\chi)\right],$$

where χ is the time lag between electric field and alternating optical shift, A is the direct optical shift, and B is the amplitude of the alternating shift. The value of the direct photo-current $i_1=i_0\sin^2 A/2$ is determined with a galvanometer, while the wave form of the alternating photo-current is observed on the screen of a cathode-ray oscilloscope. The length of the Kerr cell is chosen in such a way that the direct birefringence causes an optical shift of several multiples of 2π. If a voltage is applied to the plates of the Kerr cell, it is observed on the screen of the cathode-ray oscilloscope that the amplitude of the alternating photo-current passes through a series of maxima and minima. If the voltage is turned off, the vibrating birefringence disappears without passing through the maxima and minima.

This observation leads to the conclusion that there are two different relaxation times τ_A and τ_B for direct and alternating effect where $\tau_A>\tau_B$. If the field is turned on, the vibrating optical shift assumes its final value within a time τ_B. The direct optical shift, on the other hand, reaches its ultimate value within a considerably longer time τ_A. During this time τ_A the alternating photo-current passes through a minimum whenever $A=n\pi$. If, on the other hand, $A=n\pi/2$, the amplitude of the vibrating photo-current assumes a maximum value. If the voltage is turned off, the alternating optical shift B vanishes within a time τ_B, and the vibrating component of the photo-current decays within the same time.

In this connection it is interesting to note that Errera, Oostveen, and Sack[4] found similar effects in connection with investigations of the magneto-optical effect of colloidal solutions of iron oxide. They found that the relaxation time of the birefringence was different, depending on whether direct or alternating magnetic fields were used for the measurements. They came to the conclusion that the birefringence in direct fields had a relaxation time between 1/2 and 1/50 sec., while the relaxation time in alternating fields was about 1/1000 sec. In accord with Majorana's[6] earlier experiments they found that the birefringence was positive for small magnetic fields and negative for high fields.

A comparison of their findings with measurements reported here suggests that the electro-optical effect of bentonite sols is the electric analogy to the magnetic birefringence of iron oxide sols. Errera, Oostveen, and Sack tried to explain their measurements by the presence of particles of different diameters; they assumed that the larger particles were permanent dipoles, whereas the orientation of the small particles was caused by their anisotropy. The fact that the bentonite sols were centrifuged monodisperse solutions seems to exclude such an explanation of the present problem.

The presence of two relaxation times might find a possible explanation by the assumption that only part of the birefringence is caused by orientation of the anisotropic micelles, and that another part of the double refraction is caused by interaction between the micelles. Such an interaction effect was treated by Havelock[17] who considered the birefringence caused by dipole-dipole interaction between spherical particles. If part of the steady birefringence were caused by such an interaction effect, one might be able to explain the presence of two relaxation times.

ACKNOWLEDGMENT

In conclusion the author would like to express his deep appreciation for the many helpful discussions with Professor Mueller who suggested this problem and under whose direction this work was carried out. At the same time the author wishes to thank Professor Hauser and Dr. Le Beau who were kind enough to furnish the monodisperse bentonite sols, and whose cooperation made this work possible. The writer is further indebted to Mr. G. J. Yevick for help in the preparation of the manuscript.

[17] T. H. Havelock, Proc. Roy. Soc. London **77**, 178 (1906).

An Application of Spectrographic Methods to Chemical Concentrations of Trace Elements in Iron and Steel Analysis

R. A. WOLFE AND R. G. FOWLER
Department of Engineering Research, University of Michigan, Ann Arbor, Michigan
(Received November 9, 1944)

INTRODUCTION

THE experimental methods discussed in this paper were developed for the Simonds Saw and Steel Company of Lockport, New York, under the auspices of the Department of Engineering Research at the University of Michigan. The investigation was undertaken to develop a method for obtaining acceptable analyses of elements present in iron alloys well below the usual limits of detection by either spectrographic or chemical procedures. Analyses involving trace elements have become important to metallurgists in recent years because some alloys have been shown to be affected in both chemical and physical properties by the presence of even minute amounts of certain elements. The problem of alloys with unwanted elements has been aggravated during the war by the scarcity of new raw materials and the use of much larger amounts of scrap. While spectrographic methods are being used at present for many routine production problems, there still remains the definite need for analyses below the usual analytical ranges now investigated. The combined spectrographic-chemical techniques discussed below were designed to answer additional questions put to the laboratory analyst by control personnel. In addition, the work described illustrates a general method of attack that should become more important in commercial applications as the field of spectrographic analysis is extended.

GENERAL PROCEDURES

Spectrographic methods of obtaining an absolute determination of the composition of a given substance in solutions are well established. Solutions similar in composition to the unknown are prepared containing known additions of the desired impurities. The solutions are then compared with the unknown by reference to a suitable internal control element present in, or added to, both standards and sample in constant proportion to the whole.

To insure the absolute accuracy of the methods, it is necessary to use pure base materials in which the impurities to be analyzed for are not detectable spectrographically. In some cases even undetectable traces may influence the analyses, especially if the sample is to be concentrated before analysis and if by means of chemical concentration methods it can be shown that these traces are of the same order of magnitude as the impurities in the unknown sample. In such cases special techniques are necessary.

Precision by this method depends greatly on the nature of the principal constituent of the solutions employed. Some compounds give erratic spectrographic results, while others are found to be entirely reliable.

If iron alloys are to be analyzed by solution methods, the most obvious procedure would be to dissolve the samples in some acid such as hydrochloric, nitric, sulphuric, or perchloric, and use the salt so formed as a coating on suitable electrodes in an a.c. or d.c. arc. Of the possible compounds formed, ferric chloride is the only salt which is sufficiently water-soluble to permit electrodes to be loaded with an amount of material large enough for analytical purposes. Reduction in the amount of salt used on the electrodes effects a simultaneous reduction in the absolute amounts of impurities available to be detected, and usually a corresponding reduction in sensitivity.

A second method, more complicated but offering better chances of detecting small traces of elements, consists of the removal of the iron from the prepared solutions and the subsequent concentration of the remaining material. Different methods of accomplishing the iron separation are known. The usefulness of a particular method depends on the elements it is desired to isolate. After the iron has been removed, a considerable

concentration of the remaining salts may then be effected. Concentration methods must be resorted to whenever the quantity of the desired element is undetectable spectrographically, as is frequently found to be the case for regular solution analyses.

ANALYSIS OF FERRIC CHLORIDE SOLUTIONS

Pure ferric chloride solutions, at a concentration of 10 percent $FeCl_3$, were prepared by extracting the salt with isopropyl ether from a solution of C.P. $FeCl_3$ in 20 percent hydrochloric acid. The purified ferric chloride was recovered by washing the isopropyl ether with redistilled water. No impurities could be detected in this material spectrographically with a 5-ampere a.c.

TABLE I. Detection of impurities in ferric chloride.

Element	Concentration detectable
Aluminum	0.0001 percent
Antimony	0.02 percent
Arsenic	0.01 percent
Bismuth	0.005 percent
Cadmium	0.0001 percent
Chromium	0.005 percent
Cobalt	0.001 percent
Copper	0.00005 percent
Magnesium	0.0001 percent
Manganese	0.0001 percent
Molybdenum	0.002 percent
Nickel	0.001 percent
Titanium	0.02 percent
Vanadium	0.03 percent
Zinc	0.001 percent

arc. The requisite amounts of impurity elements were then added to portions of this purified solution to give a series of suitable standards extending from 0.0001 percent to 0.1 percent. Table I gives the limits of detection of fifteen elements in an aqueous solution of ferric chloride under the following spectrographic conditions: Ends of high purity graphite electrodes were coated with polystyrene and dried; 0.15 ml of a 10 percent solution of the $FeCl_3$ was then placed on the end of each prepared electrode and thoroughly dried; the samples were run in a 2200-volt a.c. arc at 5 amperes; electrode separation 1.5 mm.

Attempts to use ferric chloride as a medium for the analysis of iron alloys proved unsuccessful. Errors of the order of 100 percent were encountered between successive spectra. The use of chlorides of the alkaline earths and of the alkali metals as buffers failed to improve the results sufficiently to give satisfactory results. The reason for the failure of the ferric chloride to give reproducible results is not known but it seemed to center around the high volatility of the salt, most of which was dispelled from the electrodes in less than a minute after starting the arc. Smaller arc currents did not improve the results and stability-time curves showed that the discharge was never at equilibrium under any of the test conditions.

The use of other iron salts such as the nitrate and sulphate likewise proved unsuccessful. The lower solubility of these salts in water and their higher volatility in the arc made results too uncertain to depend upon.

CONCENTRATION METHOD

Sodium salts, particularly sodium hydroxide, have been found reliable as a base material for analysis.[1] Element interferences have been absent up to at least 0.1 percent for all constituents tested and the precision of analysis has been very satisfactory. To obtain high precision it is necessary that the test elements be in true solution. Large amounts of iron would not remain in a basic solution, and in precipitating, would carry down unpredictable amounts of other elements with a consequent loss of precision. It thus became necessary to remove the iron from the test solutions before dissolving the remaining elements in sodium hydroxide.

According to the reports of reputable laboratories in regard to the ethyl ether extraction method, the degree of separation attained for the desired group of elements is given in Table II.[2] Since it would have taken considerable time to make a rigorous quantitative check on the isopropyl ether separation for all the elements given in Table II, a confirmatory test only was made. For test purposes a 20 percent solution of hydrochloric acid was used to dissolve ferric chloride containing 0.1 percent of each of the following elements: Al, As, Sb, Bi, Cd, Cr, Co,

[1] Duffendack and Wolfe, Ind. Eng. Chem. **10**, 161 (1938). E. A. Boettner, Ind. Eng. Chem. **13**, 861 (1941).

[2] *Sampling and Analysis of Carbon and Alloy Steels*, Chemists of the United States Steel Corporation (Reinhold Publishing Corporation, New York, 1938), p. 67.

Cu, Mg, Mn, Mo, Ni, Ti, V, Zn, and Sn. The mixture of ferric chloride and impurities was treated with isopropyl ether. The ferric chloride was then recovered from the ether by shaking the solution with distilled water. The aqueous solution of ferric chloride was then examined spectrographically for traces of the added impurities, with the results shown in Table III.

In the actual analysis it was found that molybdenum was not completely lost to the isopropyl ether as would have been expected from a study of Table II, in which ethyl ether was used. In fact, a comparison between the accepted procedure under which the solubilities in Table II were reported and the special procedure which was finally adopted showed that only one-half as much molybdenum was extracted by the isopropyl ether as by the ethyl ether.

Analysis of the elements arsenic, antimony, and tin by the isopropyl ether extraction method was complicated by the solubility of the chlorides of these elements in the ether and by the loss of any remaining quantities when the residual chlorides were evaporated to reduce the volume of liquid. Further study is needed for these elements to determine a correct chemical concentration method before proceeding with the final spectrographic determinations.

PREPARATION OF SAMPLES

Samples were used varying in weight with the range in which the analysis was expected to fall. The largest sample used in this investigation was 10 grams. The sample was dissolved in 20 percent HCl obtained as the constant boiling point mixture when commercial C.P. HCl was purified by distillation. The iron was oxidized in the presence of an excess of HCl with about 10 ml of distilled 30 percent hydrogen peroxide. The oxidized solution was then evaporated nearly to dryness. The concentrated mixture was taken up with a volume of distilled 20 percent HCl equal to about five times the original weight of the sample.

TABLE II. Solubilities of metallic chlorides in ethyl ether.

Soluble	Mo
Partially soluble	As, P, Sb, Sn
Very slightly soluble	Cu, Zn, V
Totally insoluble	Al, Ag, Bi, Ca, Cd, Cr, Co, Pb, Mn, Ni, Ti, W, Zr

The iron was removed by extraction with distilled isopropyl ether in a separatory funnel. While one separation with an excess of ether can be used if speed is necessary, it is usually more desirable to use fractions of the ether volume and repeated separations. About three times as large a volume of ether (in ml) as sample (in grams) was required, and usually three extractions were made with equal volumes.

TABLE III. Residual amounts of impurities remaining in recovered ferric chloride.

1.	Mn, Cr	0.0000 percent
2.	Cd	0.0006 percent
	Mg	0.00003 percent
	Al	0.00009 percent
	Cu	0.00015 percent
	Zn	0.0015 percent
	Ni	0.003 percent
	Co	0.003 percent
3.	Bismuth could not be checked accurately because of the complications of neighboring iron lines.	
4.	As, Sb, V, and Ti could not be checked because of their poor sensitivity in ferric chloride.	

The final concentration of the acid solution of the residual elements can be carried to whatever extent is desired. Obviously, the final concentration necessary will depend upon the amount of the element in the unknown sample and the percentage of that element which can be accurately determined spectrographically in a sodium hydroxide mixture.

For this investigation the volume was reduced below 5 ml, 0.3 gram of twice recrystallized tartaric acid, dissolved in a small amount of water, was added followed by 5 ml of 50 percent sodium hydroxide containing 0.002 gram of germanium per ml as internal standard. The tartaric acid was added to the solution to prevent the precipitation of the remaining iron. The total volume was then made up to 10 ml with redistilled water.

PREPARATION OF STANDARDS

Standards were prepared quickly and accurately by dilution methods. For this purpose a set of auxiliary water solutions was made up containing standard amounts of the test elements. Concentrations of one gram of metal per hundred ml of water were found suitable. A

series of standard solutions was then conveniently made by taking the proper portions of the auxiliary solutions to give the desired concentrations.

It is, of course, necessary to use only pure materials for all standards. Acids should be distilled as many times as are required by the smallness of the quantities of impurities to be detected. Tartaric acid may be purified by repeated recrystallizations from distilled water solutions. This purification was found to be necessary because ordinary C.P. tartaric acid contains many metallic impurities. Sodium hydroxide of high purity may be obtained commercially.

Pure hydrogen peroxide was obtained by vacuum distillation of the 30 percent "Superoxol" grade. The operation was conducted in a completely automatic still. While the use of hydrogen peroxide as oxidant proved to be extremely successful and seemed to increase the retention of certain elements, the use of nitric acid should also prove satisfactory in most cases.

PREPARATION OF THE INTERNAL STANDARD

Germanium was chosen as the internal standard because it was not present in the steels to be analyzed and because it has proved to be a good internal standard in sodium hydroxide mixtures. Germanium dioxide was dissolved by fusion in platinum in about twice its weight of sodium carbonate. The sodium germanate formed was then dissolved in a minimum amount of water by gently warming water in the crucible itself. The resulting solution was then made up to a predetermined volume of internal standard and caustic solution by the addition of the appropriate amount of 50 percent sodium hydroxide. A value of 0.002 g of Ge/ml was found to be satisfactory.

TABLE IV.

Element	Co	Ni	Ti	Zn	Ag	Cu	V
Peroxide I							
Isopropyl ether	0.026	0.013	0.008	0.018	0.0011	0.027	0.018
Peroxide II							
Isopropyl ether	0.023	0.020	0.009	0.018	0.0016	0.055	0.017
Nitric acid							
Ethyl ether	0.034	0.013	0.009	0.008	0.0013	0.047	0.020
	Mo	Al	Bi	Cr	Mn	Mg	Pb
Peroxide I							
Isopropyl ether	0.020	0.008	0.028	0.025	0.034	0.0005	0.013
Peroxide II							
Isopropyl ether	0.015	0.011	0.025	0.027	0.039	0.0002	0.013
Nitric acid							
Ethyl ether	0.009	0.010	0.015	0.019	0.038	—	0.019

SPECTROGRAPHIC TECHNIQUE

Purified graphite electrodes are required but they need not be coated with a protecting layer before applying the test material. They were loaded with one drop of the test solution by means of a pipette. The total amount of material placed on the electrodes is not critical and this easy method of loading is adequate. The electrodes were then dried at a temperature of 100°C.

The source used was a 2200-volt a.c. arc at 4 amperes with an electrode separation of 1½ mm. The arc was run for 30 seconds before the record exposure was begun, to allow the salts to be fused and dried thoroughly. After this 30-second preburn, the record exposure may be extended for several minutes if necessary to bring out weak lines.

Standard photographic procedures were used and attention need only be drawn to the choice of photographic plate. Because of the extremely wide wave-length separations between reference and test element lines, it was necessary to use a plate, the contrast of which remained relatively constant over a broad wave-length region. Eastman Polychrome, Kodak 33, and Cramer Contrast plates were found suitable.

The working curves were set up in the usual manner by plotting the logarithm of the ratio of the intensity of the line of the test element to the intensity of the reference line, against the logarithm of the concentration of the unknown. Most of the curves, except in cases where the line intensities were very low, resulted in straight lines. The working curves were found to be stable if the conditions under which they were set up were carefully repeated. It should be necessary to check against standard samples only rarely.

PRECISION AND ACCURACY

The precision of the analyses of the sodium hydroxide solution alone for the test elements has been found to be of the order of 5 percent average error. It is difficult to assess the accu-

TABLE V. Comparison of spectrographic and chemical determinations of steel samples.

	1	2	3	4
Chromium				
Spectrographic	0.006	0.015	0.025	0.03
Chemical	—	0.016	0.027	0.034
Added	—	0.010	0.020	0.04
Manganese				
Spectrographic	0.015	0.03	0.035	0.06
Chemical	—	0.035	0.047	0.065
Added	—	0.02	0.04	0.08
Molybdenum				
Spectrographic	—	0.009	0.015	0.030
Chemical	—	0.012	0.021	0.046
Added	—	0.01	0.02	0.04
Nickel				
Spectrographic	0.013	0.019	0.035	0.05
Chemical	—	0.024	0.033	0.07
Added	—	0.01	0.02	0.04
Vanadium				
Spectrographic	0.0003	0.009	0.018	0.018
Chemical	—	0.014	0.025	0.024
Added	—	0.01	0.02	0.04

racy of the entire procedure, including the extraction, concentration, and spectrographic analysis. Table IV shows results obtained on the same sample using isopropyl ether and ethyl ether for extracting the ferric chloride and nitric acid and hydrogen peroxide for changing the iron to the ferric form. Two determinations were made with hydrogen peroxide and isopropyl ether to show the precision of the recommended method. The ethyl ether-nitric acid procedure is shown for comparison only.

The results shown in Table IV indicate that discrepancies of from 10 to 50 percent exist, depending on the element being considered. Some of these discrepancies, particularly that of molybdenum, can be attributed to the difference between methods and may represent a loss of the material during extraction. By referring to Table II, it is at once evident that the greatest variation might occur in the elements Mo, Cu, V, and Zn, and this fact is in general borne out by the analyses in Table IV.

Best results on repeat tests were found with the peroxide-isopropyl ether method. By keeping all experimental conditions as nearly constant as possible, it is believed that samples can be analyzed with errors no greater than 25 percent of the amount of the element present.

AGREEMENT OF RESULTS WITH CHEMICAL DETERMINATIONS

Four samples of steel were analyzed by the peroxide-isopropyl ether method, and five elements in three of the samples checked by chemical methods at the laboratories of the Simonds Saw and Steel Company at Lockport, New York. The results are shown in Table V. Sample No. 1 was pure commercial iron and the indicated small amounts of various elements were added to it to give samples Nos. 2, 3, and 4. The alloys were made in an induction furnace and precautions were taken to prevent loss of the added impurities.

The agreement between spectrographic and chemical results is quite good. On the whole, the

TABLE VI. Self-consistency of spectrographic determinations.

Element	1	2	3	4
Aluminum				
Analyzed	0.006	0.01	0.011	0.04
Corrected	0.004	0.009	0.014	0.033
Bismuth				
Analyzed	—	0.01	0.025	0.040
Corrected	—	0.013	0.020	0.026
Chromium				
Analyzed	0.006	0.015	0.025	0.030
Corrected	0.009	0.013	0.022	0.030
Cobalt				
Analyzed	0.005	0.034	0.040	0.09
Corrected	0.004	0.036	0.057	0.092
Lead				
Analyzed	0.0015	0.007	0.012	0.032
Corrected	0.0015	0.008	0.012	0.032
Manganese				
Analyzed	0.015	0.03	0.035	0.06
Corrected	0.018	0.026	0.037	0.048
Molybdenum				
Analyzed	—	0.009	0.015	0.030
Corrected	—	0.008	0.014	0.033
Nickel				
Analyzed	0.013	0.019	0.035	0.05
Corrected	0.011	0.019	0.034	0.054
Silver				
Analyzed	0.0004	0.0009	0.0013	0.003
Corrected	0.0004	0.0012	0.0015	0.0031
Titanium				
Analyzed	—	0.003	0.009	0.01
Corrected	—	0.0025	0.009	0.018
Vanadium				
Analyzed	—	0.009	0.018	0.018
Corrected	0.0003	0.010	0.018	0.018

spectrographic analyses are probably nearer the truth in the low ranges, while the chemical analyses are better in the higher ranges. One test of the consistency of the spectrographic results should be pointed out. The pure iron, as shown in Table V, sample No. 1, has 0.006 percent residual chromium present and the analysis showed that samples Nos. 2 and 3 contain 0.005 percent Cr in excess of the amount added.

CONSISTENCY OF THE SPECTROGRAPHIC VALUES

The four samples were checked against each other for consistency by testing them as metallic pins in an a.c. arc and determining the intensity ratios of the various elements to appropriate iron lines. When these results were plotted against concentration, as determined by extraction methods, and the empirical law that all analytical curves are linear was applied, it was found that in most cases the departure of the analyses from the curve was small. A correction could then be made on the analysis to give the value required to restore all points to a straight line. Values in the higher concentration ranges were selected for this test so that spectrographic results could be compared with chemical analyses. Since very little correction was found to be necessary in the higher analytical ranges, it seems reasonable to assume that results obtained by extraction-spectrographic methods in the lower concentrations would be equally valid. The results show further that the percentages of the various elements lost to the isopropyl ether must be a proportionate amount of the element present; otherwise, a straight line working curve would not be found. A comparison is given in Table VI between the analyzed spectrographic values and the corrected spectrographic values to show the consistency of the measurements. The analyzed spectrographic values were determined from working curves made from synthetic mixtures.

The line pairs selected for making the analyses given in Table VI are tabulated in Table VII.

TABLE VII. Line pairs used in analysis.

Aluminum	3082.16A	Germanium	3039.09A
Bismuth	3067.73	Germanium	3039.09
Cadmium	2288.03	Germanium	2754.59
Chromium	2843.25	Germanium	2754.59
Cobalt	3453.51	Germanium	3269.50
Copper	3273.97	Germanium	3269.50
Lead	2833.07	Germanium	2754.59
Magnesium	2802.71	Germanium	2754.59
Manganese	2798.27	Germanium	2754.59
Molybdenum	3170.35	Germanium	3269.50
Nickel	3414.77	Germanium	3269.50
Silver	3280.67	Germanium	3269.50
Tin	2839.99	Germanium	2754.59
Titanium	3349.04	Germanium	3269.50
Vanadium	3185.41	Germanium	3269.50
Zinc	3345.0	Germanium	3269.50

CONCLUSIONS

In general, the method is limited in speed and accuracy largely by the chemical techniques involved. However, since these techniques would normally be required anyway in trace analyses, the method can be said to be more rapid than the chemical procedure, since it does not require individual separations of the elements. The accuracy reported is sufficiently good to be of use in many important research and control applications. It should be pointed out that many difficult chemical techniques might be greatly aided by using some modification of the suggested procedure. If chemical concentration can be employed on a trace analysis problem, it is highly probable that the spectrograph can be used to complete the process.

ACKNOWLEDGMENT

The experimental work on this problem was very ably carried out by Mr. Wayne W. Shapiro.

Letter to the Editor

Note on Low Refraction and Dispersion of Beryllium Fluoride Glass

ALEXIS G. PINCUS
Research Division, American Optical Company, Southbridge, Massachusetts
August 24, 1944

THE glass-forming properties of beryllium fluoride were first predicted and demonstrated by V. M. Goldschmidt.[1] He also predicted that it would have a low refractive index and a low dispersion, but apparently never measured them, stating only that the refractive index was lower than that of water.

Biltz and Rahlfs[2] measured the density of a beryllium fluoride glass at 25°C as 1.986 g/cc.

K. H. Sun[3] in May, 1943 stated that: "No refractive index value has been given in the literature. Safford (private communication), however, stated that it is lower than that of methyl alcohol (1.3)."

Recently we have discovered a source of beryllium fluoride glass of quality suitable for refractivity measurements. The Brush Beryllium Company[4] makes this glass by melting crystallized ammonium beryllium fluoride in graphite pots and pouring into a similar pot. It is available in chunks several inches across, which contain occluded carbon, blowholes, and bubbles. Its purity is given as:

BeF_2	about 99.9 percent
Fe	0.01 to 0.03 percent
Al	0.06 percent
Si	0.02 percent.

As long as they are kept in a tight container, the chunks stay highly transparent and bright. But the glass is extremely hygroscopic. On exposure to the atmosphere on a humid day, the surfaces immediately become sticky, and this sticky portion will flow down into a puddle.

By special techniques, prisms of the glass have been polished. The surfaces have to be protected against atmospheric moisture, as even in a desiccator over sulfuric acid the polish deteriorates. For spectrometer measurements it was necessary to protect the surfaces with polished glass plates 1 mm thick.

Table I gives the refractive indices obtained for the *D*, *F*, and *C* lines by two observers.

TABLE I. Refractive properties of beryllium fluoride glass.

	Observer 1	Observer 2	Averages
N_C	1.2739_2	1.2739_2	1.2739_2
N_D	1.2747_5	1.2747_4	1.2747_5
N_F	1.2765_0	1.2764_6	1.2764_8
$N_F - N_C$	0.00258	0.00256	0.00257
N_H	106.3	107.3	106.9

ACKNOWLEDGMENTS

Mr. Carl G. Silverberg surfaced the prism and maintained its polish. Spectrometer measurements were taken by Dr. A. Estelle Glancy and Mr. J. K. Davis.

[1] V. M. Goldschmidt, Vid. Akad. Oslo I. Matem.-Natur. **1926**, No. 8 (1927).
[2] W. Biltz and E. Rahlfs, Zeits. f. anorg. allgem. Chemie **166**, 374 (1927).
[3] K. H. Sun, Glass Ind. **24**, 218 (1943).
[4] Information in this paragraph obtained by private communications from Brush Beryllium Company.